Cognitive Neuropsychology
(in press)

Interactivity and Continuity in Normal and Aphasic Language Production

Wheeler Ruml, Alfonso Caramazza
Harvard University

Rita Capasso
Fondazione Santa Lucia IRCCS, Roma

Gabriele Miceli
Università Cattolica

Abstract

Current research in cognitive modeling has assumed that the interactive nature of processing during language production has been supported by fits to the behavior of brain-damaged patients. In this paper, several previously-proposed theories of word production, all based on the interactive spreading-activation theory of Dell (1986), are evaluated in the context of picture naming. Using a new corpus of data from 50 Italian aphasic patients, we find that patient patterns do not seem to demand an interactive theory. We also explore the corollary ideas of continuity between normal and random performance, and global damage in aphasia, and we find that they are incompatible with our data. We argue that it is very difficult for a modeling study to yield strong constraints on cognitive theory. Although valuable, such efforts currently do not contribute evidence for interactive processing in language production.

Wheeler Ruml, Division of Engineering and Applied Sciences, Harvard University; Alfonso Caramazza, Cognitive Neuropsychology Laboratory, Harvard University; Rita Capasso, Fondazione Santa Lucia IRCCS, Roma; Gabriele Miceli, Istituto di Neurologia, Università Cattolica.

Thanks to Michele Miozzo, Gary Dell, the Harvard Cognitive Neuropsychology Laboratory, and Stuart Shieber for useful suggestions and stimulating discussions regarding this research. Support was provided in part by National Science Foundation grants IRI-9350192 and IRI-9618848, National Institutes of Health grant DC 04542, a grant from MIUR, and a short-term mobility grant from CNR.

Please address correspondence concerning this article to Wheeler Ruml, ruml@parc.com, Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304 USA.

The extent to which the various representations and processes involved in a particular cognitive behavior interact is a fundamental question. In language production, for example, interaction between phoneme representations and word representations is believed to be responsible for phenomena such as the mixed error effect. Dell and Reich (1981) found that, when an incorrect word is produced that is semantically related to the intended word, it is more likely to be phonologically related to the intended word than one might expect by chance (see also Shallice and McGill (1978)). One explanation for this phenomenon involves a feedback hypothesis in which active phoneme representations excite word representations before the word to be uttered has been chosen.[1] The feedback makes it more likely that a word phonologically related to the intended target word will be selected in the event of an error.

One approach that researchers have taken in order to evaluate such theories of language processing is to implement a computational simulation of the theory and then evaluate the model's ability to match human performance (Dell, 1986; Roelofs, 1992; Coltheart, Curtis, Atkins, & Haller, 1993). In addition to using data from ordinary participants, much work has focused on data from brain-damaged patients (Plaut & Shallice, 1993; Harley & MacAndrew, 1995; Rapp & Goldrick, 2000) By including damaged processing components in the computer simulation, models can attempt to account for the patient data.

This approach was used by Dell, Schwartz, Martin, Saffran, and Gagnon (1997), who presented the picture-naming performance of 21 aphasic patients along with the output of a computer simulation of brain-damaged naming. Dell et al.'s theory includes a broad hypothesis about the interactive structure of lexical processing. Dell et al. suggest that the various representations in the lexicon communicate during word production not only in the forward direction (from semantic to lexical to phonological information) but simultaneously in the backward direction as well. This is reflected in their simulation by the spreading of activation along bi-directional connections between representations. Their proposal also includes two other major claims. One of these is the *continuity thesis*, which relates the performance of normal and aphasic participants. Motivated by the idea of explaining impaired behavior as a modification of normal processing, continuity means that "more severe aphasic patients have an error pattern that is closer to the error opportunities afforded by the lexicon, whereas less severe aphasic patients have a pattern that is similar to the normal pattern" (p. 820). The final claim is the *globality assumption*, which states that, in fluent aphasia, "the damage involves all the levels in the lexical network" (p. 814). Although it is no longer used as a working hypothesis, (Ruml & Caramazza, 2000; Foygel & Dell, 2000), we will see later in this paper how this simple assumption, along with the continuity thesis, is closely related to the interactive nature of the model.

Although some recent studies have illustrated discrepancies between Dell et al.'s interactive model of lexical retrieval and patient data, they have all depended on relatively small numbers of patients (Ruml & Caramazza, 2000; Ruml, Caramazza, Shelton, & Chialant, 2000; Caramazza, Papagno, & Ruml, 2000; Cuetos, Aguado, & Caramazza, 2000). It can be difficult to tell from such studies whether the discrepant patients represent important test cases or merely outliers in the data. In this paper, we will evaluate several models of picture-naming using a new corpus of data from 50 Italian aphasic patients, all of whom

[1] Although they will not be the focus of this paper, other mechanisms besides feedback have been proposed to explain mixed errors, such as the recurrent attractor networks of Plaut and Shallice (1993).

have been tested using Dell et al.'s methodology and whose responses have all been scored using similar criteria. This is more than twice the number reported by Dell et al. (1997) and three times the number reported by Ruml et al. (2000). As an additional benefit, the patients we report here are native Italian speakers. This provides an opportunity to test the ability of Dell et al.'s model, and other proposals based on it, to generalize beyond the English-speaking patient data that have been previously reported. We adapt the model to Italian using a novel lexicon construction procedure and show that it fits control data better than its English counterpart. We then test its ability to account for our patient data. In addition to the original aphasic damage assumptions of Dell et al. (1997), we also test more recent proposals by Foygel and Dell (2000) and Rapp and Goldrick (2000). We find that, despite the good fit of the base model to control data, each specific proposal for modeling aphasic naming fails to capture important features of the Italian patient data. Furthermore, a non-interactive version of Foygel and Dell's proposal fits the data as well as the original interactive one. We argue that these results remove aphasic patient data from the set of results supporting interactive spreading-activation models of lexical access.

Interactivity itself is a broad, unfalsifiable concept. But as we will discuss, the continuity thesis and the globality assumption are two specific, closely related claims that are open to test. Having a large group of patients allows us to begin assessing claims about the distribution of patient error profiles, such as the continuity thesis. With a small number of patients, the absence of particular patterns has little significance. With our larger corpus, we find that the variety of error patterns patients can exhibit is incompatible with continuity. Instead of behaving randomly, we argue that impaired patients can exhibit dissociations, in which only one component of lexical access is damaged, leading to impaired but highly non-random behavior.

Many details must be specified before a simulation can be applied to patient data. After a brief introduction to Dell et al.'s model of lexical access and to our patient data, we will discuss the construction of an Italian analogue of Dell et al.'s lexical network, which will allow us to use the model with our Italian speakers. Then we will proceed to the evaluation of the model.

Background

First, we will discuss Dell et al.'s theory and computational model of lexical access. After seeing the model of normal performance, we will outline three proposed methods for relating the base model to aphasic patients.

Interactive Theories of Lexical Access

The base model of normal performance is based on notions taken from the lexical access theory of Dell (1986). The model is connectionist in style; its structure is indicated in Figure 1. The undamaged system is postulated to consist of three levels of interacting representations: semantic, lexical, and phonological. Each level interacts only with the adjacent levels. The representations in each level are instantiated as real-valued activation values. These values are updated during processing according to the activation levels of neighboring representations, whose influence is controlled by a *connection weight* parameter, and the decay of the original activation, controlled by a *decay* parameter. Random noise is

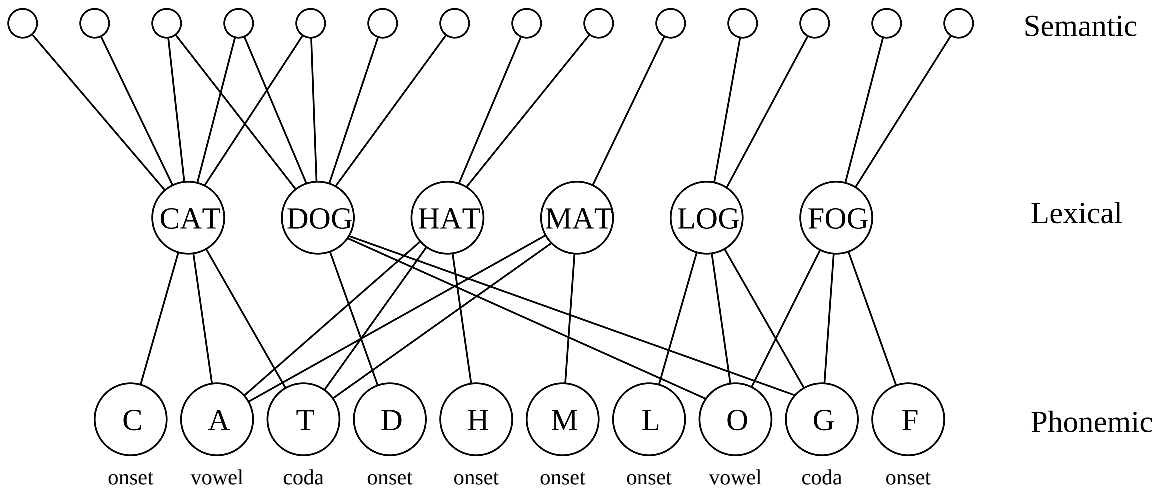

Figure 1. The structure of part of Dell et al.'s model.

also added to each node independently (see Dell et al., 1997 for details). All connections in the network are bidirectional, thereby implementing the theoretical idea of interaction between adjacent levels.

Lexical access is simulated in the model by raising the activation level of the semantic representations associated with the target word to a predefined constant value, simulating the spread of activation throughout the network for eight time steps, and then raising the activation level of the most active lexical node. This corresponds to the notion of lexical selection. Activation is further propagated for another eight time steps, after which the most active onset, vowel, and coda phonemes are chosen as the output of the model. Activation is cascading in the sense that phonological representations become active well before the lexical selection step. This allows them to feed backward and influence choices at higher levels.

Because of the contribution of noise to the activation levels, the selected lexical representation does not necessarily correspond to the target word, and even if it does, the selected phonological representations are not necessarily those associated with the selected lexical node. This allows the model to simulate naming errors. Dell et al. divide their model's errors into five types: formal, which are responses that are words that sound like the depicted object (e.g., 'hat' for 'cat'), semantic (e.g., 'dog' for 'cat'), mixed, which are both formally and semantically related (e.g., 'rat' for 'cat'), unrelated (e.g., 'log' for 'cat'), and nonword (e.g., 'dat' for 'cat'). By simulating many lexical access trials, one can accumulate an estimate of the probabilities of the various types of responses (such as correct, formal error, or nonword) and determine whether the model represents a mechanism sufficient to generate a response distribution measured from a human experimental participant.

Constructing the Lexical Network

We will discuss in some detail the methodology that Dell et al. followed in constructing their model's network because a similar methodology will be used later in this paper to construct a version for Italian speakers. To reduce the computational burden of simulation, they chose a lexical layer of six words and distinguished a particular word as always playing the role of the target. The phonological layer follows immediately, being determined by the lexical nodes and English pronunciation. To construct the semantic layer, Dell et al. stipulated that all lexical representations are associated with ten semantic features and that semantically related words share three of their features.

To provide some assurance that this small network captures some of the relevant properties of English, Dell et al. measured its error opportunities, that is, the probability that a response of each possible type would result from selecting a random phoneme in the model at each syllabic position. This represents the model's behavior when it is generating random phonologically-legal strings of phonemes. (Legality is enforced by the CVC structure of the model.) They compare this distribution to an estimate of random error opportunities for English-speaking aphasic patients. For patients, these opportunities represent a simple model of complete breakdown of the lexical access system. They were estimated in a two-step process using stimuli from picture-naming experiments. First, following Dell and Reich (1981) and Best (1996), Dell et al. estimated the probability that randomly replacing the first phoneme in a target word results in an English word. (To ensure phonological legality, this was done only with words having a vowel in second position.) This also establishes the nonword probability. Dell et al. estimated a .8 probability of a nonword error and .2 chance of a word error. Then, to decide how the word errors should be apportioned among error categories, they used the analysis of Martin, Dell, Saffran, and Schwartz (1994). Martin et al. used target-word error pairs obtained from aphasic patients in picture-naming tasks to generate random re-pairings. Scoring these and multiplying by .2 yielded estimates of .09 formal errors, .01 semantic, .004 mixed, and .10 unrelated. By postulating two separate small network structures, only one of which contains a possible mixed error, and specifying that one is used 90% of the time while the other (with the mixed error) is used the remaining 10% of the time, Dell et al. were able to obtain error opportunities for their simulation model that accorded to their satisfaction with the estimates for English. Furthermore, they showed that the model can generate an error distribution rather similar to that of control participants in a picture-naming task.

Theories of Damage in Aphasia

Since the model's error distribution depends on the simulation parameters, changing the parameters can cause the model to emit errors following a different distribution. Dell et al. (1997) propose that the damage to the lexical access system in aphasia be modeled as changes to their simulation's *connection weight* and *activation decay* parameters. By changing these parameters throughout the network, this model of brain damage attempts to explain various patterns of errors by uniform damage to all parts of the system simultaneously, thereby implementing Dell et al.'s globality assumption. By finding, for each patient, the values for *connection* and *decay* that allow the model to best match that patient's error distribution (χ^2 test), one can assess the model's fit. Automated regression procedures for finding good parameter values have been proposed by Ruml and Caramazza (2000) and Foygel and Dell (2000). Although it has been argued (Ruml & Caramazza, 2000; Caramazza et al., 2000; Rapp & Goldrick, 2000) that a general explanation of aphasia in terms of global damage is precluded by the reports of patients whose impairment seems limited to only one subsystem of lexical access, the original global damage model was an influential proposal and continues to enjoy prominence in the literature, so we include it in our tests here.

In response to the evidence against global damage, Foygel and Dell (2000) proposed an alternative model of brain damage. Under this model, the values of *connection* used between the semantic and lexical layers and between the lexical and phonological layers can each vary from the usual level. This allows errors of lexical selection, while preserving correct selection of the corresponding phonemes. Although concerns have been raised regarding the performance of this model (Ruml et al., 2000), Foygel and Dell claim it "is at least as good as the [global] model in reproducing the data" (p. 35). It is certainly one of the strongest current models of aphasic lexical access.

Finally, we will also briefly consider the model of aphasia proposed by Rapp and Goldrick (2000). The underlying model is similar to Dell et al.'s, using spreading activation in a multi-level network, but it lacks the feedback from lexical nodes to the semantic layer. A fourth 'conceptual' layer lies above the semantic features. Damage is introduced by adding noise at the conceptual, lexical or phonological levels. Because feedback is not a universal principle in this model, it is known as the "restricted interaction" account. They found this model to provide closer matches to their patients under a wider variety of parameter settings than Dell et al.'s model.

Data from Italian Aphasic Patients

With an understanding of Dell et al.'s model of word production, we can now prepare to test its ability to account for human behavior. First, we present an overview of the data that we collected from 50 Italian fluent aphasic patients. Then we will turn to the modifications we made to the model in order to apply it to the Italian data.

Patient Selection and Testing

Aphasic subjects were included in our study independent of etiology, lesion site and speech rate (with certain exceptions noted below). Dysarthric subjects were excluded if articulatory problems prevented unequivocal scoring of their responses. There were two additional disqualifying conditions: 1. Unstable medical or neuropsychological status, resulting either in the inability to carry out the cognitive evaluation in two weeks or in marked fluctuations in performance from one session to another; 2. Diffuse (i.e., non-focal) brain damage, as revealed by CT-scan or MRI, or diffuse cognitive impairment, as detected by the Mini Mental Status Exam (Folstein, Folstein, & McHugh, 1975) and by the Mental Deterioration Battery (Carlesimo, Caltagirone, & Gainotti, 1996).

Our sample includes 33 males and 17 females, of whom 48 are right-handers and 2 are corrected left-handers. Of these, 42 suffered from cerebrovascular accidents (CVAs), 3 from Primary Progressive Aphasia (PPA) and 5 from Herpes Virus Encephalitis (HVE). Details are provided in Appendix A.

Our criteria for patient inclusion differ slightly from those used by Dell et al., who only considered subjects with left posterior CVAs resulting in fluent aphasias. We also included subjects with reduced speech rate following left anterior CVAs, subjects with PPA and with HVE. Cases of HVE were included, even though their lesion distribution differs from that observed in CVAs, because HVE typically has a sudden onset, followed by gradual recovery, just like CVAs. Primary progressive aphasia has an insidious onset and a slow progression, just like dementing illnesses, but behavioral deficits are restricted to the language domain even at very late stages, and do not affect other cognitive abilities. Of 3 cases of PPA in our sample, GBU had non-fluent speech without dysarthria, whereas BCO and AME presented with semantic dementia. They were tested at least 2 years post-onset, as proposed by Weintraub, Rubin, and Mesulam (1990) to prevent the incorrect classification of subjects with initial Alzheimer's disease as cases of PPA.

All of the subjects included in this study were submitted to the Batteria per l'Analisi dei Deficit Afasici (Miceli, Laudanna, Burani, & Capasso, 1994), a 36-test battery that analyzes performance on tasks exploring the sublexical, lexical-semantic and grammatical levels of language. Some of the subjects included in this study have been reported in detail elsewhere (FS: Miceli & Caramazza, 1988; SF: Miceli, Giustolisi, & Caramazza, 1991; CLB: Miceli & Caramazza, 1993; ECA: Miceli, Capasso, & Caramazza, 1994, 1999; GMA and RBO: Miceli, Amitrano, Capasso, & Caramazza, 1996; WMA: Miceli, Benvegnù, Capasso, & Caramazza, 1997; PGE and GIM: Miceli & Capasso, 1997; APA: Miceli et al., 2000; IFA and AS: Caramazza, Chialant, Capasso, & Miceli, 2000; SVE: Miceli & Capasso, 2001). All of the subjects completed the experimental tasks in the same order: naming was evaluated first, then repetition, then comprehension. Subjects were tested over a period of no more than 20 days, with each testing session lasting up to 60 minutes. No more than two sessions per day were held and no more than one task was administered on each day.

A group of 21 cognitively unimpaired subjects served as control for our brain-damaged group. These subjects undertook the same naming and repetition tasks as the brain-damaged group. Subjects were informed that their results would be used as a match against the performance of persons with language disorders resulting from brain damage, but were not aware of the purposes of the study. Only one selection criterion was used: since all the aphasics in our sample had 8 or more years of formal education, only controls with at least an eighth grade education were considered. They were each tested in one session, which usually lasted between 15 and 25 minutes.

Patient Picture-Naming

A summary of the patients' repetition and comprehension performance is provided in Appendix A. We focus here on the test of picture-naming performance, as that is the central task Dell et al.'s model is intended to capture.

Stimuli and Scoring Criteria.

Naming abilities were investigated in the normal controls and in the aphasic subjects by means of a task that comprises 128 pictured nouns, belonging to 11 semantic categories (animals, body parts, fruits and vegetables, food, professions , kitchenware, clothing, tools, furniture, means of transportation, musical instruments).[2] Each picture was presented on

[2]Exceptions: Subject SDI completed three administrations of the naming task. Subjects PGE, SF, EMA,

a separate sheet. The subject was instructed to respond with one word whenever he could. The task was administered without time limits; however, if a subject failed to provide a response within one minute (this never happened with normals, and only infrequently with aphasics), or showed obvious signs of distress at the inability to produce the correct response, the next stimulus was presented.

We followed the scoring methodology used by Dell et al. (1997), with certain exceptions dictated by features of the Italian language. Only the first patient response was scored. Responses were classified as correct, semantic error, formal error, mixed error, unrelated error, nonword, or other, according to the following criteria:

Correct: The expected word or a synonym (*macchina*, car → *automobile*). By analogy with Dell et al., inflectional errors were also scored as correct responses. In Italian, these errors result almost always in the substitution of the final vowel (*gamba*, leg → *gambe*, legs).[3] Inflectional errors accounted for 93 of 7,777 total responses in brain-damaged subjects (1.2%).

Semantic: An incorrect word that was a category superordinate, coordinate, subordinate, or associate. Single-word incorrect responses resulting in a verb conceptually related to the target (*ago*, needle → *cucire*, to sew; *forbice*, scissors → *taglia*, (it) cuts) were also scored as semantic errors. By contrast, multi-word utterances focused on a verb (*coltello* , knife → *per tagliare*, for cutting; *biscotto*, cookie → *si mangia*, you eat it) were scored as circumlocutory responses and included in the *Other* category.

Formal: Dell et al. scored as formal errors incorrect word responses that started or ended with the same phoneme as the target, or shared with the target more than one phoneme in any position, or one phoneme in the same within-syllable position. In Italian, such criteria would result in a disproportionately high number of formal errors due to two main factors:

1. When evaluating a response for formal similarity, Dell et al. excluded unstressed vowels from the phoneme count. Italian phonology does not include the 'schwa' sound and, even though a vowel does not necessarily carry word stress (e.g., in the word /'tavolo/, /a/ carries word stress, but the two /o/'s are full vowels), there are no unstressed vowels in the language. Consequently, according to Dell el al.'s criterion, all the vowels in an incorrect response should be counted when scoring for formal similarity between target and error. Furthermore, the vast majority of Italian words are polysyllabic, and contain on average more stressed vowels than English words. These facts greatly increase the probability of scoring an incorrect response as a formal or mixed error (and decrease to a comparable extent the probability to score an incorrect response as a semantic or unrelated error).

2. In English, almost any consonant and vowel can be the last phoneme in a word. By contrast, all Italian words are inflected and, with the exception of loan words, end

GMA, and RBO were asked to name other stimuli as well. Because these subjects performed comparably across administrations of the same task, or across various naming tasks, only their overall performance is discussed here. Subjects CLB and AS were tested on a subset of the stimuli. They are included in this report because the evaluation of their naming skills is sufficiently extensive as to provide reliable results.

[3]The stimuli did not include any nouns with irregular plural endings (e.g., *bue*, ox → *buoi*, oxen).

with one of four vowel sounds (*a*, *e*, *i*, *o*). There are very few words ending with *u*, but these are of very low frequency, with the exception of a few function words. As a consequence, an Italian speaker who responds to a picture by randomly selecting a lexical entry will produce an inflected word, whose final vowel is very likely to increase formal similarity by sheer chance.[4]

To circumvent these difficulties, we designed an alternative criterion for formal similarity that is slightly more conservative. Incorrect responses were considered to be formally related to the target if they shared one-third of the phonemes with the target, irrespective of their sequence (*scrivania*, desk → *pavimento*, floor) or if they shared the two initial phonemes (*telefono*, telephone → *tennista*, tennis player; *stringa*, string → *stelle*, stars), two phonemes in the first syllable (*francobollo*, stamp → *fastidi*, nuisances; *nave*, ship → *ancora*, anchor), or three phonemes in the same sequence in any position within the stimulus (*ginestre* , brooms → *straccio*, mop; *campana*, bell → *ricamo*, embroidery; *asino*, donkey → *stringhe*, shoe strings). The inflectional vowel was ignored. We will present the patients' naming performance below using the original criterion as well as this new one.

Mixed: Incorrect word responses bearing both a formal and a semantic relationship (as defined above) to the target. Examples of these errors include *gomito*, elbow → *fronte*, forehead (shared one-third of phonemes); *rapa*, beet → *radice*, root (shared first two phonemes).

Unrelated: Incorrect single-word responses that bore neither a formal nor a semantic relationship to the target. Following Dell et al., each response were scored with respect to the picture for which it had been produced and thus a perseverated response was often classified as unrelated. There was no attempt to keep score of the distance between the stimulus that induced the first perseverated response and the subsequent perseveration(s). This is a minor problem with our subject sample, as these errors occurred to a significant extent only in subject FDP.

Nonword: Any response that did not correspond to an entry in the Italian vocabulary, independent of the number of phonemes it shared with the stimulus (i.e., independent of whether it was formally related or unrelated to the target). We included in this category both random phoneme strings that obeyed the phonotactic constraints of the Italian language and pseudo-morphological neologisms. The latter type of error occurred very infrequently (35 of 7,777 total responses, 0.4%), and resulted from the illegal combination of a root morpheme with an inappropriate inflection (*baffi*, mustache, m.pl. → "baffa", pseudofeminine) or with an inappropriate derivation (*libreria*, bookshelf → "portalibri", book container). Semantic substitutions that deviated from the target at the segmental level and resulted in a non-existing word were also included in this category (*naso*, nose → "pento", pseudoword presumably standing for *mento*, chin). 103 responses were of this type (1.3%), with 56 from

[4] This is also true in the case of a subject who responds to a picture by means of an unrelated, random sequence of phonemes that obey the phonotactic constraints of Italian. Some responses will result in an 'inflected' target, whose final vowel might contribute to formal similarity by sheer chance (*elefante*, elephant → /kro'pibe/).

just four subjects (FBI, EGI, MIO, WMA), whose lesions extended into the anterior portions of the left hemisphere.

Other: The following error types were grouped in this category:

1. Non-responses. The subject failed to respond in one minute, or comments that cannot be construed as attempts at naming the target (*ananas*, pineapple → "I have one at home").
2. Multi-word responses. Definitions and circumlocutions produced as attempts to describe the target (*ambulanza*, ambulance → *porta via le persone malate*, it takes away people who are sick).
3. Part/Whole responses. Incorrect responses in which the subject provides the name of only a portion of the stimulus (*automobile* car → *ruote*, wheels).
4. Visual errors. The subject provides the name of an object that is visually similar, but otherwise unrelated to the target (*gesso*, chalk → *riga*, ruler).

We will use *strict scoring* to refer to the use of the tighter formal criterion (with its concomitant effects on semantic and mixed errors) and *loose scoring* to refer to the use of Dell et al.'s original criterion.

Overview of Performance.

The picture-naming performance of each patient is listed in Appendix A. An overview of these data is shown in Figure 2 using a variant of the popular 'schematic plot' format of Tukey (1977). For each response type, a vertically-oriented symbol shows how the frequency of that response type varied across patients. For each symbol, the middle white box shows the range of the middle 50% of the patients, the horizontal line across the box marks the median, and the vertical whiskers extend to the minimum and maximum. Outlying values further than four quartiles from the median are shown as tiny circles, and the gray stripe indicates a 95% confidence interval around the mean.[5] The figure shows the distributions over each response type for the English speakers reported by Dell et al. (1997) and for our Italian patients. In the semantic, formal, mixed, and unrelated categories, we show the distributions over Italian patients using both the relatively loose criterion for formal relatedness of Dell et al. and our stricter requirements.

The figure indicates that the Italian patients are in some respects similar to those reported by Dell et al. Most responses are either correct, nonword errors, or unscorable 'other' responses. Both sets contain patients who name correctly on more then 90% of trials, although the Italian set contains more patients with poor performance (both the minimum, median, and quartiles are lower, and the mean may be lower). The distributions over nonword frequencies are also similar. But the figure also reveals some notable differences between our patients and those reported by Dell et al.

First, the Italian set contains many more patients with high frequencies of 'other' responses. Fully half of the Italian patients have 'other' rates greater than 10%. Such patients are quite common among fluent aphasics and explaining their retrieval failures is an important area of current research. Their scarcity among Dell et al.'s population may

[5]Confidence intervals were calculated by assuming that the frequencies are values drawn from a normal distribution.

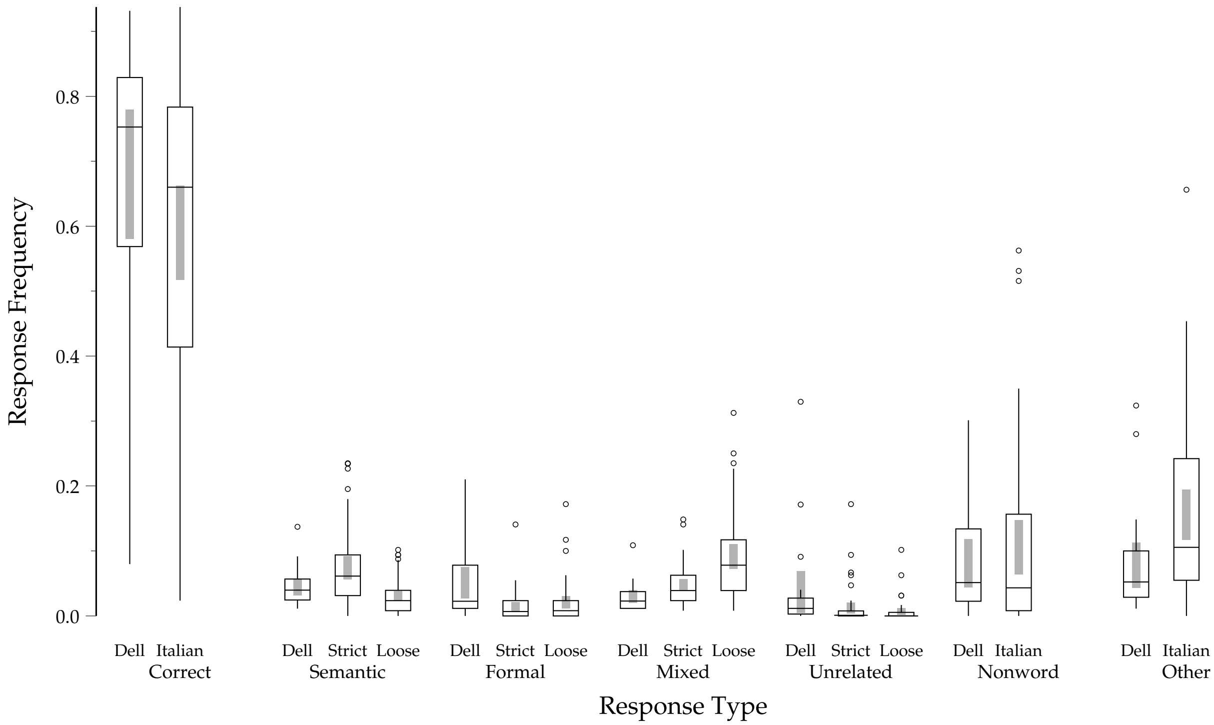

Figure 2. The distributions of response frequencies among our patients (labeled Italian, Strict, and Loose) and those reported by Dell et al. The boxes represents the middle 50% of each distribution, lines mark the medians, gray stripes give 95% confidence intervals on the mean, and circles represent outliers.

have resulted from concerns regarding the fit of Dell et al.'s computational implementation, which does not simulate complete failure of retrieval.

Second, the Italian patients made fewer purely formal errors and more mixed errors. Using the looser scoring criterion, the 25th percentile of the mixed error rates in the Italian data is near the 75th percentile of the English data. Rates of 10% or higher seem unrealistically inflated in comparison to the rates of the English-speaking subjects for whom Dell et al.'s model was originally designed, and all data presented in later sections of this paper will use the stricter criterion unless otherwise noted. Even using the stricter scoring, 75% of the Italian patients made more mixed errors than the median Dell et al. patient, and 25% (about 12) made more than all but one of Dell et al.'s patients (about 4%). Not surprisingly, those responses that were mixed errors became semantic errors under the stricter formal criterion. With the stricter criterion, the Italian patients make many more semantic errors than their English counterparts. Twenty-five percent of the Italians made more semantic errors than all but one outlier among Dell et al.'s patients. Several Italian patients made more than 15% semantic errors. The difference in criteria has little effect on the rates of formal or unrelated errors. Both types were less frequent among the Italian patients, with 75% of them making fewer such errors than the median Dell et al. patient.

These differences do not seem attributable to significant differences between Italian and English speakers in general. Table 1 reports the grouped performance of 21 normal

Table 1: Naming Performance of Normal Participants.

Source	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Other
Italian (strict)	.978	.010	.000	.009	.000	.000	.003
Italian (loose)		.003	.000	.016	.000		
Dell et al.	.969	.012	.001	.009	.003	.000	.007

Italian control participants along with the data for English controls presented by Dell et al. (1997). Italian normals made significantly more correct responses, rather than fewer, and significantly fewer 'other' responses, rather than more. The Italian patient set just seems to contain more patients with poorer performance that Dell et al.'s set. Although the Italian controls made significantly fewer unrelated responses, differences in the other response types were not significant and do not explain the dearth of formal errors and the frequency of semantic and mixed errors among the patients.

The high frequency of semantic errors may be due to subtle sample bias. As our laboratory is known in the Rome aphasia community for interest in semantic disorders, we may see an unusually high proportion of patients with semantic damage. One indication of this in our data is the unusually high frequency of patients with Herpes Virus Encephalitis (10%), which is associated with semantic damage. The low incidence of purely formal errors may be due to features of Italian. For instance, the language has few short content words, making it difficult to obtain formal errors under phoneme substitution. (We will see further evidence of this later, in Table 4, when we calculate error opportunities in Italian.) Because the computational models we will be evaluating are all intended to account for a broad range of aphasias, including semantic disorders, any bias should not affect our conclusions.

Modifying the Simulation for Italian

Before we can attempt to model these data using Dell et al.'s simulation of naming and a theory of damage, we must make the appropriate modifications to the model's underlying lexical network. As we discussed earlier, Dell et al. determined the structure of their model's network by matching their estimate of the error opportunities in English. Thus, in order to use their model with Italian-speaking patients, we must estimate the error opportunities of Italian and construct a new network that matches them.

Error Opportunities in Italian

To estimate the error opportunities in Italian, we followed the procedure and scoring criteria used by Dell et al.. We began by estimating the probability of a nonword error. We phonetically transcribed the pronunciations of 128 stimuli which were used with most of the patients in the picture-naming task. We separated the initial phonemes, leaving us with a group of initial phonemes and a group of headless stems. Then we repeatedly paired a random initial phoneme with a random final sequence. We examined 250 such random combinations and determined how many were actual Italian words. 36 were words, 35 were phonologically illegal, and 9 were phonologically legal but had initial syllables which never occur in Italian. This left 170 nonword combinations. We ignored the phonologically

Table 2: Error opportunities in Italian, estimated according to the methodology of Dell et al. (1997). Results are shown using both criteria of formal relatedness.

	Scoring	
	Strict	Loose
Correct	0	0
Semantic	.011	.006
Formal	.039	.112
Mixed	.006	.011
Unrelated	.119	.047
Nonword	.825	.825

questionable sequences, and concluded that the opportunity for a random response to yield a non-word was $1 - (36/206) = 0.825$.

Continuing to follow Dell et al., we estimated frequencies for the other response types by dividing up the remaining probability (.175) using estimates derived from patient errors. We compiled a list of 563 word errors made by patients on our picture-naming task. We separated the patient responses from the target words of those trials, and then repeatedly selected a random target and a random word response. We created 200 random pairings and then scored them using each of the two scoring criteria discussed earlier. Our standard criterion resulted in the following word error frequencies: .065 semantic, .22 formal, .035 mixed, and .68 unrelated. (Using the looser criterion, we obtained .035 semantic, .63 formal, .065 mixed, and .27 unrelated.) Table 2 gives the final error opportunity estimates, combining the probabilities of nonword and the various word errors.

Phonological Overlap in Italian

In addition to error opportunities, we also want our network to exhibit the phonological overlap characteristics suggested by Rapp and Goldrick (2000). They proposed a measure of phonological overlap and applied it to a set of English words. The overlap of two words is computed as the number of phonemes they share (regardless of order and including duplicates), divided by the sum of the number of phonemes in each word. Rapp and Goldrick measured the average overlap both between words in the same general semantic category (such as furniture, or transportation) and between words in different semantic categories. Using eight different categories, each containing at least 19 words, they obtained estimates of .175 and .148, respectively.

To obtain analogous results for Italian, we merely translated the words used by Rapp and Goldrick and recalculated the mean overlaps. Thirteen of their 217 words were not used: 7 of the 31 birds did not have a common name in Italian and 6 other words had Italian translations that were the same as the translations of other words in the list (e.g., slacks and pants are both *pantaloni* in Italian). We obtained estimates of .245 for the mean within-category overlap, and .232 for the mean across-category overlap. Unlike Rapp and Goldrick's finding for English, this difference was not significant ($t(7) = 1.1$, $p \approx .15$ for one-tailed test). We hypothesize that the higher values for Italian are a reflection of its use of fewer vowels than English.

A Lexical Network for Italian

Given our estimates of error opportunities and phonological overlap in Italian, we then attempted to construct a lexical network that exhibited similar properties. We followed Dell et al.'s methodology of computing the error opportunities of a model lexicon as the probability of each response type when a random onset, vowel, and coda are selected. For measuring the phonological overlap, we followed Rapp and Goldrick in considering only pairs in which one of the words was the target. (This makes sense because the model's lexicon is intended to reflect properties of a typical word, including its phonological relations to other semantically related words. The overlap within other semantic categories, and indeed the existence of semantic relations between the other words of the lexicon, are ignored by the model.) Paralleling our scoring procedure, we did not include inflectional endings in the lexicon, since they can be generated by a simple regular process. This also allowed us to use the same processing formalism as Dell et al., including its assumption of CVC structure.

Rather than attempting to construct a lexicon with the desired characteristics by hand, we used an automated search procedure (described in Appendix B). This allowed us to test many more possibilities that we could have manually. As Rapp and Goldrick did, we constructed a single lexical network containing all possible types of error responses, rather than having two networks containing different types of responses among which simulation trials are allotted with certain frequencies (as in Dell et al.'s work). This avoids any inherent limit on the frequency of any given response type.

The lexicon we used in our experiments has 24 words, and is summarized in Table 3. Since it is intended to correspond to a typical lexical neighborhood rather than specific words, we have denoted the pronunciations using numbers rather than particular phonemes. The words marked 'semantic' and 'mixed' are considered to be semantically related to each other and the target. (Like Rapp and Goldrick we feel there is little purpose in finding a set of actual words that have this pattern of semantic and phonological overlap.) The resulting lexical network has 4 onset nodes, 4 vowel nodes, 9 coda nodes, and 231 semantic nodes, connected by 312 bidirectional links.

This network structure yields error opportunities that are similar to those we calculated for Italian. The distribution is shown in Table 4. Using either the RMSD or X^2 similarity metrics, we can quantify the match of the network's opportunities to the desired ones, and compare it to the match of previously proposed models for English. As shown in the table, the Italian model matches the desired opportunities better than previous English models. The network also exhibits a higher phonological overlap within the target's semantic category than to other words, as desired. Table 5 compares the Italian network's match to those of other proposed networks. It seems to match as well as Rapp and Goldrick's model. (Dell et al.'s was not designed with the overlap criterion in mind.)

We also tested the resulting model's fit to the performance of normal non-aphasic Italian speakers. Because the model always selects 3 phonemes as its output, it never issues a circumlocution or fails to respond, as humans sometimes do. To prevent this from forcing a poor fit, we follow Dell et al. and ignore 'other' responses when measuring the fit of the model. Table 6 presents fits to normal control data of both Dell et al.'s English model and our Italian lexicon. For both models, the *connection* and *decay* parameters were tuned using our automated fitting procedure. Although we found that many values worked well

Table 3: The Italian model's lexicon. Phonemes are represented by numbers.

		Pronunciation		
	Scoring	Onset	Vowel	Coda
1.	Target	1	2	3
2.	Semantic	4	5	6
3.	"	4	7	8
4.	Formal	1	2	8
5.	"	1	2	9
6.	"	1	5	12
7.	"	1	7	3
8.	"	1	11	3
9.	"	10	2	3
10.	Mixed	1	2	13
11.	Unrelated	4	5	12
12.	"	4	7	10
13.	"	4	7	14
14.	"	4	11	13
15.	"	6	5	13
16.	"	6	5	14
17.	"	6	5	15
18.	"	6	7	6
19.	"	6	7	8
20.	"	6	11	10
21.	"	10	5	8
22.	"	10	5	10
23.	"	10	5	14
24.	"	10	7	6

Table 4: Match of the Italian model to desired error opportunities.

	Response Pattern						Fit	
Source	Corr.	Sem.	For.	Mixed	Unrel.	Non.	RMSD	X^2
Italian opportunities	0	.011	.039	.006	.119	.825		
Italian network	.007	.014	.042	.007	.097	.833	.010	.010
English opportunities	0	.01	.09	.004	.1	.8		
Rapp and Goldrick	.009	.018	.098	.009	.125	.741	.027	.018
Dell et al.	.042	.042	.079	.004	.083	.75	.031	.065

Note: RMSD stands for root mean squared deviation.

Table 5: Match of the Italian model to desired phonological overlap.

	Overlap	
Source	Within	Across
Italian estimate	.245	.232
Italian network	.222	.183
English estimate	.175	.148
Rapp and Goldrick	.222	.173
Dell et al.	.033	.32

Table 6: Fit of Italian Model (decay, connection) to Italian Control Participants.

	Naming response						Fit		
Data source	Corr.	Sem.	For.	Mixed	Unrel.	Non.	RMSD	X^2	p
Italian Controls	2,628	28	0	23	0	0			
conn .027, *dec* .50	9,805	107	0	88	0	0	.000	.02	1.00
English Controls	10,094	120	6	90	28	5			
conn .0806, *dec* .55	9,748	164	0	85	0	3	.002	41.7	.000

for the Italian model, we attempted to choose final values similar to those used by Dell et al. The Italian model's fit to the control data is excellent, with $p > .9999$ even with thousands of samples from both the controls and the model.[6] Following Dell et al., we also verified that the simulation generated a mixed error effect at lexical selection: more than 1/3 of the semantically-related word nodes that were selected were the mixed word. These results give us confidence that our modifications to the underlying network should allow us to test the two models of aphasic naming using our data from Italian patients.

Modeling the Italian Patients

With an Italian version of the lexical network in hand, we are ready to evaluate the ability of the various interactive spreading-activation models to account for the aphasic patient data. We begin by testing the simplest theory of damage: Dell et al.'s original globality assumption.

The Global Damage Model

Under the global damage assumption of Dell et al. (1997), connection strength and decay rate are modified throughout the network. We evaluated this model's fit to the Italian patients in two ways. The first way was to vary each of the two parameters throughout its

[6]Throughout this paper, we report significance assuming five degrees of freedom, which is conservative since often two parameters are adjusted to fit the six dimensional data.

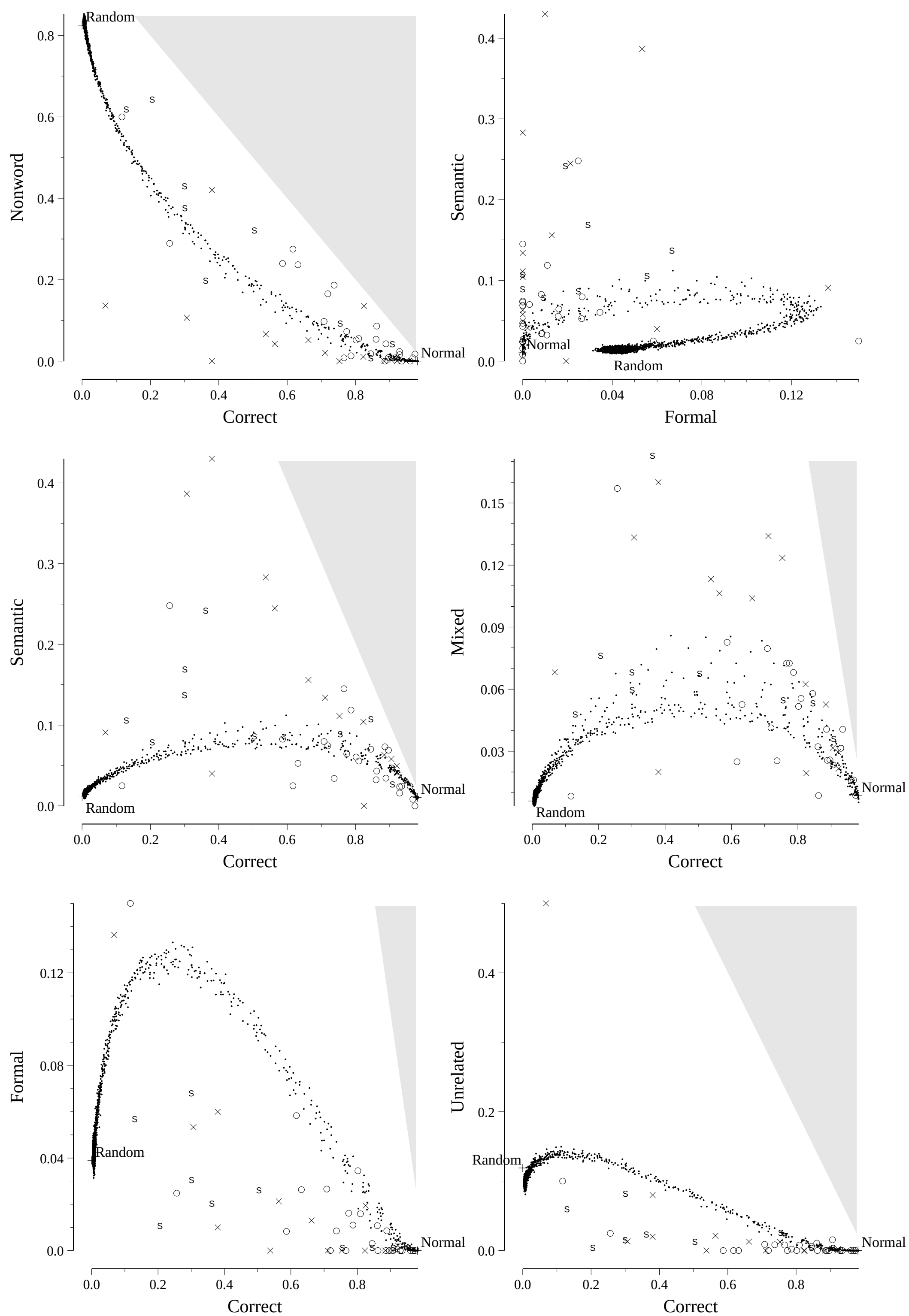

Figure 3. Possible error patterns using the global damage model of Dell et al. (1997). Patients with many 'other' responses are marked with x, those with slow speech with s, and the rest with o. Normal performance and random error opportunities are marked with pluses.

range and test the model's behavior at each combination. This gives us an understanding of the complete range of error patterns that the model can generate. Figure 3 shows the possible range of the model as tiny dots, with patients plotted as small circles, x's, and s's. Each panel shows a particular pair of response types. Each dot in a particular panel corresponds to an error pattern that was generated by the model under some setting of the parameters. The position of the dot in the panel indicates the observed frequencies of those two response types in the error pattern. The extent of the dotted regions shows the model's capabilities. To indicate the boundaries of the space where error patterns could possibly lie, areas in each panel where the two depicted response frequencies would sum to greater than one are shaded out. Evaluating the model's coverage of the patients using these panels is conservative in the sense that two points in a particular panel can visually overlap even if the error patterns they represent differ markedly in response categories not shown in that panel. (Similarly, the shaded triangles are conservative since they consider the two response categories alone—it is unlikely that error patterns would be composed of only correct and mixed responses, for example.) Other visualization methods, such as the principal components analysis of Foygel and Dell (2000), that collapse all response categories to two dimensions can give a misleading impression of the overlap between the model and the patients (Ruml et al., 2000). Patient error patterns have been normalized by ignoring 'other' responses such as circumlocutions. Because Dell et al. (1997) feel that patients who made more than 15% of their responses in the 'other' category may somehow have their response distributions improperly distorted in a way which may cause mismatch with their model, such patients are plotted using x's instead of circles. In addition, patients whose output was halting or slowed (disregarding word-finding pauses) are plotted using s's. Plus signs mark random error opportunities and the performance of normals.

The figure indicates that the global damage model allows only a very restricted range of error patterns. For example, the upper left panel indicates a strong link in the model between the level of correctness and the number of nonword responses. For 50% correctness, the model must generate about 20% nonword responses no matter how one manipulates the parameters. But many patients, including ones with few 'other' responses, lie outside the model's range. Dell, Schwartz, Martin, Saffran, and Gagnon (2000) speculated that discrepancies such as semantic overprediction in their English model might be due to overrepresentation of semantic errors in the network's error opportunities and therefore that "one could improve the fit by changing the networks' neighborhoods so that there are fewer semantic and more mixed opportunities" (p. 637). However, the Italian model has slightly more semantic and mixed opportunities than the random Italian pattern and yet, as the middle panels in the figure show, it generates too few such errors. These results indicate that Dell et al.'s model of aphasic naming does a poor job of accounting for Italian patients and that improving the underlying lexical network does not ameliorate its difficulties.

The figure also shows that the model does seem to obey the continuity thesis, as its behavior in most panels can be captured as defining a limited spectrum of patterns that stretches between normal performance and random errors. Those two response patterns govern the endpoints of the model's space of possible errors, although they alone do not determine its ability to cover intermediate states. (We will come to the question of whether the patients obey continuity later in this paper, after we have examined all three sets of model damage assumptions.)

Table 7: Summary of the fit of the global damage model, including the variance accounted for (VAF) in each response category, the mean category VAF, overall VAF, mean RMSD error, mean X^2, and number of patients with significant mismatches.

	VAF by response type						Overall VAF				
Other	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Mean	Wtd.	$\overline{\text{RMSD}}$	$\overline{X^2}$	Failures
any %	.922	.254	−.72	.525	.206	.833	.336	.796	.042	15.2	21/50 (42%)
≤ 15%	.942	.476	−.42	.585	−2.0	0.91	.074	.875	.028	9.61	9/28 (32%)

Next, we used the automated fitting procedure of Ruml and Caramazza (2000) to find the best possible fit of the model to each patient. Table 7 presents a summary of the fits. The first row includes all the Italian patients and the second excludes those who made many 'other' responses. (Recall that in both cases, the 'other' responses are not used in calculating fits, because they lie outside the scope of the model.) Several measures of fit are shown. The first six columns refer to the proportion of variance accounted for (VAF) by the model within each response category. This compares the performance of the model to merely guessing the mean value in each category. More formally,

$$\text{VAF} = 1 - \frac{\sum\limits_{patients} (observed - predicted)^2}{\sum\limits_{patients} (observed - mean)^2}$$

where *observed* represents the observed frequency of the given response type for a particular patient, *predicted* represents the model's predicted frequency of that response type for that patient, and *mean* is the mean observed frequency across all patients. Values between zero and one represent explanatory power on behalf of the model, with one representing perfect prediction. Values less than zero imply that the model performed worse than predicting the mean. Although the model accounts for substantial variance in the two largest categories, correct and nonword, it does a poor job of modeling related word errors.

The seventh column shows the mean of the six category-specific VAFs, and the eighth (labeled "Wtd") computes the VAF by pooling the data across all categories. This has the effect of weighting the categories by their variance and thus rewarding good performance in the correct and nonword columns.

Another common measure of fit is root mean squared deviation (RMSD). This is used to measure the similarity of a complete predicted naming pattern to the observed patient pattern. Formally,

$$\text{RMSD} = \sqrt{\frac{1}{6} \sum_{response\ types} (observed - predicted)^2} \ .$$

Column nine in Table 7 shows the mean of the RMSD values across the patients. Similarly, column ten gives the mean value of the well-known X^2 statistic. The final column shows the

Table 8: Summary of the fits of the connection-strength damage model to the patients (see Table 7 for notation).

	VAF by response type						Overall VAF				
Other	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Mean	Wtd.	$\overline{\text{RMSD}}$	$\overline{X^2}$	Failures
any %	.981	.271	−1.4	.136	.264	.967	.202	.857	.035	15.7	18/50 (36%)
≤ 15%	.985	.448	−.65	.272	−1.4	.984	.102	.919	.021	9.41	6/28 (21%)

number of patients whose best fit according to the model remained significantly different from their actual performance ($p < 0.05$). The model failed to match 42% of the patients.

In summary, the assumption of global damage in aphasia, combined with Dell et al.'s model of lexical access, matched the Italian patient data poorly. It seems too tightly constrained to model the variety of patient patterns. This confirms the findings of Rapp and Goldrick (2000) and Ruml and Caramazza (2000) with our larger Italian corpus. So we will turn next to the more refined model of damage proposed recently by Foygel and Dell (2000).

The Connection-Strength Model

We performed a similar evaluation using the localized damage model of Foygel and Dell (2000), in which the connection strength between the semantic and lexical layers and lexical and phonological layers are the two parameters to be varied. (*Decay* is held at its normal value.) Figure 4 shows the model's possible error patterns and their overlap with the patient data. As the top left panel shows, the model is able to generate a wide variety of rates of nonword errors at each level of correctness. (This important feature appears to have been overlooked by Foygel and Dell in their analysis of their English model, as they suggest that "the error space of the [connection-strength] model is quite similar to that of the [global damage] model", p. 200.) The triangle in the panel is defined by correct performance (bottom right) at normal parameter settings, random word errors (bottom left) as the weight from the semantic to the lexical level is reduced, and mostly nonword random performance (top left) as the weights to the phonological level are lowered, regardless of the semantic weights. Just as this allows the model to dissociate nonword responses from the overall level of correctness, it also represents a retreat from the continuity thesis of Dell et al.'s earlier work. Foygel and Dell's model does not assume that patient performance converges to the random error opportunities as severity increases.

Despite this added flexibility, the model still has trouble generating enough semantic errors (top right and middle left panels). It also generates too many formal word responses at almost every level of correctness (bottom left panel). Foygel and Dell speculated that the inability of the connection-weight damage model to generate enough mixed errors might be due to faulty error opportunities (p. 211). Since our lexicon has slightly more opportunities for mixed errors than the Italian random pattern, this is not likely to be the case.

A summary of fits to the Italian patients using the two-level connection-strength model is shown in Table 8. It seems to do slightly better than the global damage model in accounting for variance but it still fails to match many patients.

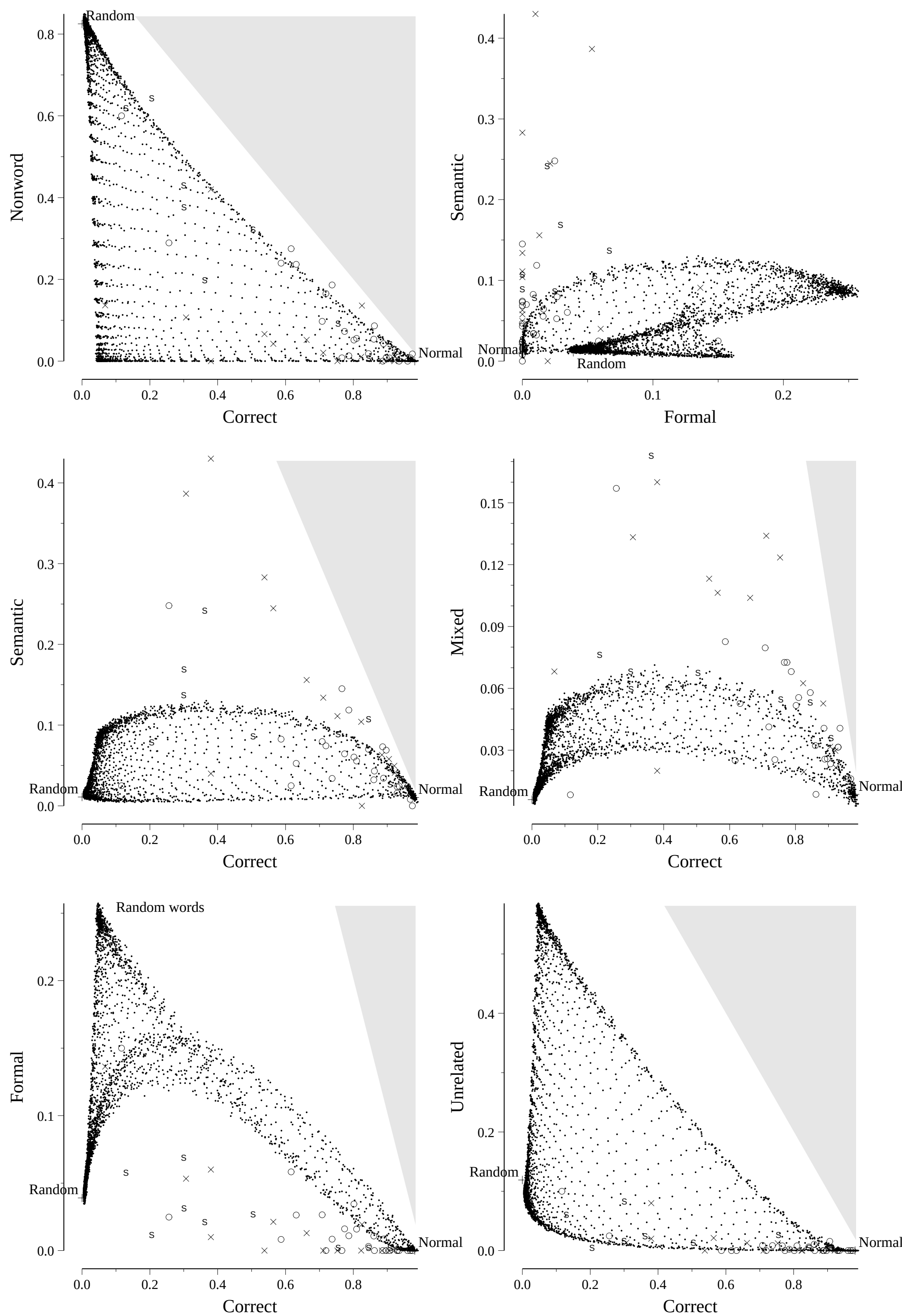

Figure 4. Possible error patterns using the connection-strength damage model of Foygel and Dell (2000).

Table 9: Summary of the fit of the repetition predictions of the connection-strength damage model (see Table 7 for notation).

	VAF by response type						Overall VAF				
Corpus	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Mean	Wtd.	$\overline{\text{RMSD}}$	$\overline{X^2}$	Failures
Italian	.585	-.04	-24.	-1.6	-1.4	0.74	-4.2	.587	.034	28.8	13/34 (38%)
Dell et al.	$-12.$	$-\infty$	-7.2	.229	$-.19$	$-14.$	$-\infty$	$-12.$	.034	54.0	5/11 (45%)

Following Dell et al. (1997), Foygel and Dell suggest using their model to predict repetition. If one assumes that the word has been correctly heard, then one can use the second half of the production process (starting at lexical selection) to model repetition. Table 9 summarizes the performance of the model when used to generate repetition predictions. Italian patients with fewer than 80 repetition trials were excluded. Performance using Dell et al.'s English lexicon and patients is shown for comparison. The model has an overall VAF of 59% on the Italian patients and fails to match on 38% of the patients. The VAF is extremely poor for Dell et al.'s patients, especially in the semantic category, in which there were always no errors in the patient data but the model sometimes predicted some incorrect responses. Overall, the model seems to capture significant variance in only the correct and nonword categories.

Understanding the Model's Behavior.

Looking again at Figure 4, we see that it shows two main problems with the model: it cannot generate enough semantic errors (middle left panel) and it generates too many formal errors (bottom left panel). The reasons for this are straightforward. Consider semantic errors. These arise almost always from misselection at the lexical level, as a result of damage to the connections from the semantic level. This connection strength damage reduces the activation level of the target to a level at which noise can often cause other words to be selected. This is essentially the only circumstance which can generate a semantic error. However, in this situation, the model must also make many formal and unrelated errors. When the target's activation level is low enough for it to be frequently confused with its semantic competitors at selection time due to noise, this means that its activation level must be very low. Background noise causes all other words to also be active. Although semantic competitors have an advantage because they receive some input activation, this advantage is very slight when connection damage has suppressed activation levels. The sheer number of unrelated and formal competitors means that one of them is likely to have received enough noise to be selected. Fundamentally, semantic errors arise as the model moves toward selecting random words at the lexical level. So semantic errors are inextricably tied to formal and unrelated errors in this model.

We have seen why formal errors must be generated when producing semantic errors. But formal errors must also be generated under damage to the connections between the lexical to phonological levels. Again, activation levels are suppressed to the point where the phonemes of the word selected at the lexical level are not necessarily those that are most active at output time. Of course, when a non-target phoneme is selected, the result is likely to be a formal or nonword error. This simple process explains why formal errors are always produced by any lesion in the connection strength model.

These processes can be seen in the patterns of Figure 4. In particular, the lower left panel clearly reflects the two ways formal errors can be produced. The model's performance envelope is defined by three points. At the lower right of the panel lies normal performance, with high correctness and few formal errors. At the top left is the error pattern produced by selecting words randomly from the lexicon (marked "Random words" in the panel). Few match the target, resulting in low correctness, but a large fraction are formally related. Finally, at the lower left, we find random performance, corresponding to choosing a random onset, vowel, and coda (marked "Random" in the panel). This results in very low correctness and a small but noticeable number of word responses that happen to be formally related to the target. The small dots showing the models performance are strung between these points, representing different parameter settings.

By following the edges of the model's performance envelope, we can trace the effects of varying each of the connection-strength parameters. As the semantic connection weight parameter is lowered from its normal value, the model's lexical nodes receive less activation. This lowers correctness and increases the number of formal errors, resulting in the band of points moving up and to the left from the normal point. A pure semantic-lexical lesion happens to correspond to the bottom edge of this band, which passes through a mass of other points and continues up and to the left. When the connections to semantic input are effectively severed, lexical nodes are selected at random, resulting in random word performance at the top left corner. Most of the other responses at this parameter setting are unrelated words, and tracing the same parameter changes on the lower right panel leads one up along the top edge toward random, most unrelated word errors.

Reducing the phonological weight from its normal value also increases the rate of formal errors. A pure lexical-phonological lesion corresponds to the top edge of the band going up from the normal point—it then curves downward, crossing the edge produced by a pure semantic-lexical lesion, ending at the bottom left at the point corresponding to random phoneme selection. Most of the other errors are nonwords, resulting from errors during phonological encoding, and the corresponding path in the upper left panel defines the top edge of the nonword triangle. At intermediate levels of damage, the number of formal errors is even greater than under random phoneme selection, because the phonemes of the target are still often selected.

As a result of these phenomena, the model is unable to generate a large rate of semantic errors, because many unrelated word errors must also occur, and it always produces many formal errors, because such errors occur under both the postulated types of damage.

The Role of Interactivity.

Surprisingly, our explanation of the model's behavior did not involve interactivity. Instead of implicating feedback from the phonological level to explain formal errors, for instance, the explanation was couched in terms of word opportunities and activation of the target's phonemes. To test our understanding, we constructed a simulation almost identical to Foygel and Dell's connection-strength model that we tested above, with the only difference being a complete lack of feedback connections. This non-interactive model matched our Italian controls well ($p > .55$), using a slightly lower normal connection strength

Table 10: Fits of the non-interactive connection-strength damage model to the patients (see Table 7 for notation).

	VAF by response type					Overall VAF					
Other	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Mean	Wtd.	$\overline{\text{RMSD}}$	$\overline{X^2}$	Failures
any %	.988	0.29	-1.4	.128	.064	.969	.174	.854	.036	15.8	19/50 (38%)
≤ 15%	.984	.515	-.76	.251	-3.0	0.96	-.17	.903	.023	9.72	7/28 (25%)

(0.025 instead of 0.027).

Its possible error patterns are shown in Figure 5. The shapes of the patterns seem indistinguishable from those in Figure 4, confirming our hypothesis that interaction does not play an important role in shaping the possible error space of Foygel and Dell's model. A summary of the non-interactive model's ability to fit patient patterns is presented in Table 10. Although the fits are very slightly worse in the unrelated and mixed categories, they are very slightly better in the correct and semantic categories, and the overall weighted VAF is essentially unchanged from the interactive model. These results show that fitting Foygel and Dell's model to aphasic patient data cannot be said to support an interactive theory of lexical access, since a purely feedforward model does just as well. Of course, both versions exhibit systematic flaws that prevent accounting for many patients.

The Restricted Interaction Model

We also tested a restricted interactivity model along the lines suggested by Rapp and Goldrick (2000). Following the structure of their model, an additional conceptual level of representation was added above the semantic feature level, with one concept node for every lexical node, connecting to the same semantic features. No feedback was allowed from the lexical to the semantic level. The default connection weight and noise parameters were adjusted to maintain a good fit to normal performance ($p > 0.98$). A three-step simulation procedure was used for each naming trial, prefacing Dell et al.'s two-step process by the selection of the most active concept and raising the activation of the corresponding semantic nodes. Damage was simulated by increasing the activation-related noise at the lexical and phonological levels independently.[7] In addition, for those patients whose comprehension was less than 95%, we allowed adjustment of the activation noise at the conceptual level to promote a more accurate fit.

Figure 6 shows the error patterns that can be generated from the model by manipulating all three tunable parameters. It can generate many more semantic word errors than the other two models of aphasia. By modeling damage as the noisy movement of node activation values at the conceptual and lexical levels, the activation of lexical semantic competitors can be increased without necessarily involving formal competitors. The model has trouble generating enough nonwords, though, and overgenerates mixed and formal responses. The increased coverage of the space of patterns led to slightly improved fits to the individual patients, as summarized in Table 11. The fits to the semantic and mixed categories are

[7]Dell et al.'s model has two types of noise: background noise, which has a constant variance, and activation noise, which is proportional to the activation of a node.

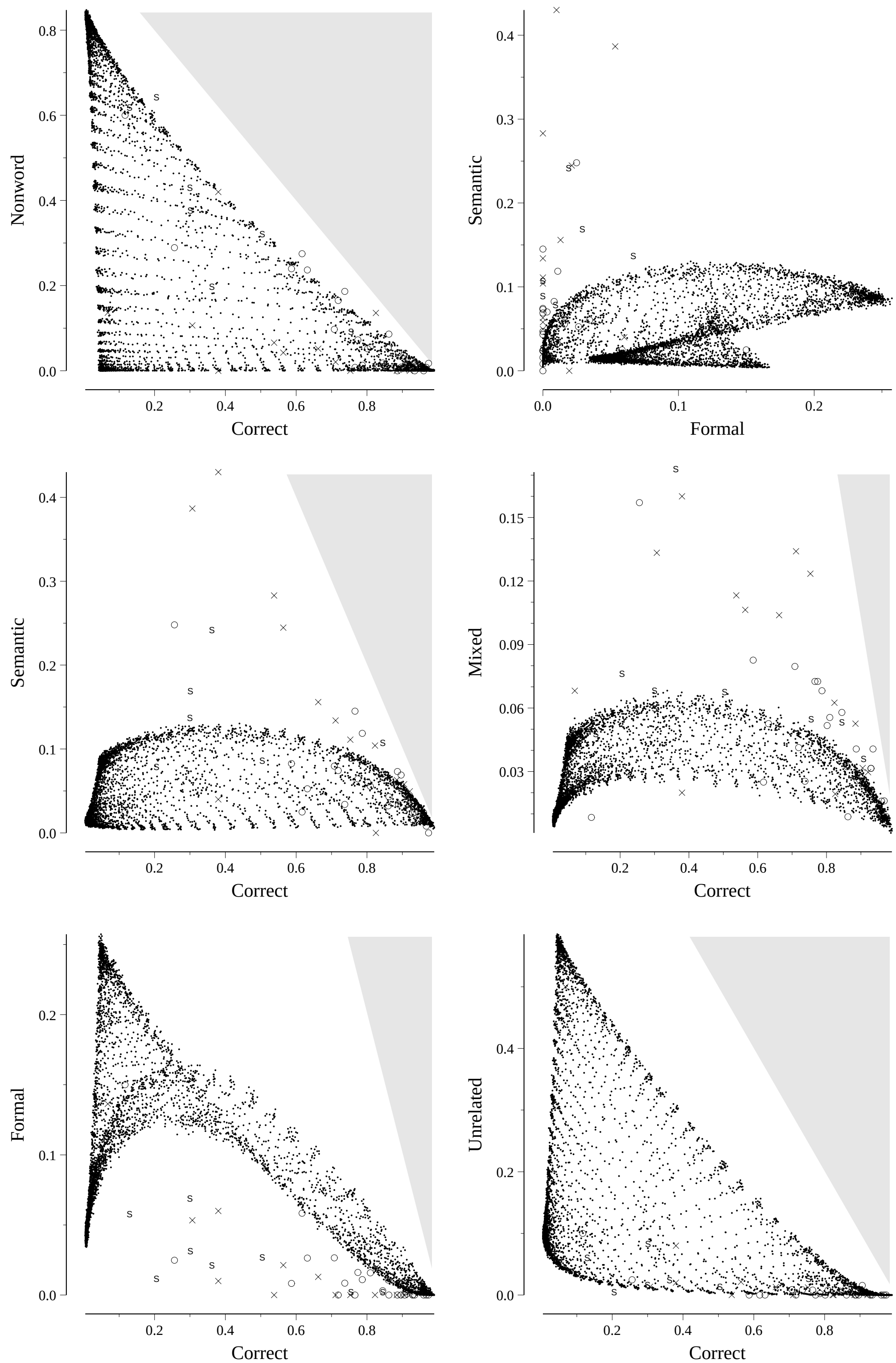

Figure 5. Possible error patterns using the connection-strength damage model of Foygel and Dell (2000) in a network with no interactivity.

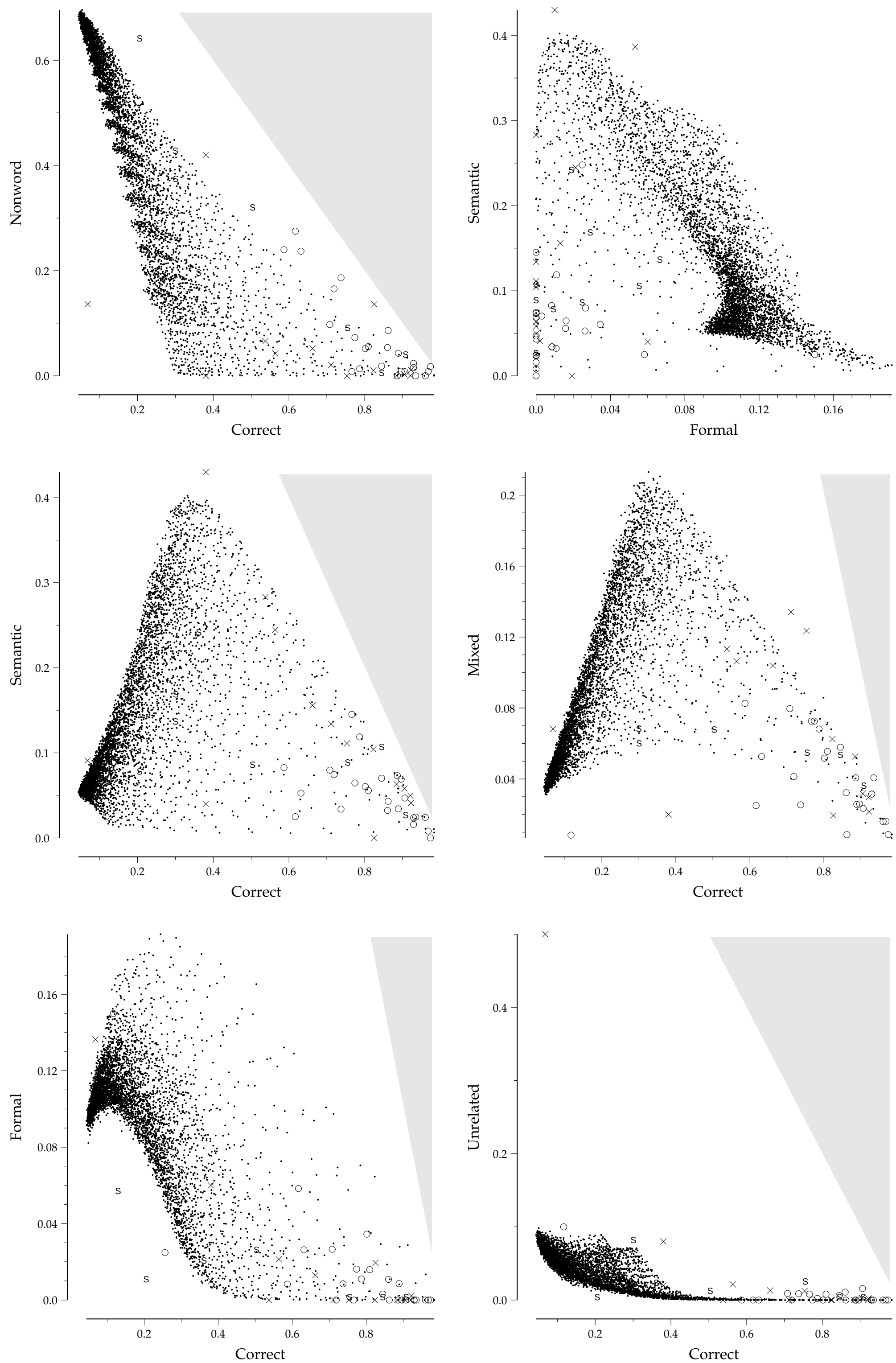

Figure 6. Possible error patterns using the restricted-interactivity model based on Rapp and Goldrick (2000).

Table 11: Fits of the limited interactivity noisy damage model to the patients (see Table 7 for notation).

	VAF by response type						Overall VAF				
Other	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Mean	Wtd.	$\overline{\text{RMSD}}$	$\overline{X^2}$	Failures
any %	.943	.922	-.433	.594	.339	.830	.532	.966	.031	11.0	14/50 (28%)
$\leq$ 15%	.916	.899	. -.563	.528	.819	.956	.592	.987	.023	10.2	7/28 (25%)

particularly good. However, the model still failed to fit 28% of the patients. And this evaluation does not take into account the fact that a third tunable parameter was adjusted for many of the patients. If we take the number of tunable parameters into account when computing significance tests for the patient fits, the number of mismatches increases to 24 out of 50 (14/28 when excluding patients with many 'other' responses).

So What?

We have now tested Dell et al.'s two-stage interactive spreading activation model using each of three models of aphasic damage. We found that no set of damage assumptions allowed the model to cover the range of observed patient behaviors, even in our two-dimensional views that ignore four response categories at a time.

But so what? Given such simplified models of human language processing, it is not surprising that they fail to fully account for complex clinical cases of aphasia. They use tiny lexicons of single-syllable CVC nouns, an impoverished semantics, and two-parameter notions of brain damage. Are we not holding these models up to an impossible standard, one that may in fact be counter-productive for research in this area? Rapp and Goldrick (2000), for instance, suggest that "a 'data-fitting' approach to evaluating the adequacy of various simulations...reflect[s] overconcern with matching aspects of the data that are not predicted by the theoretical claims under consideration" (p. 493, see also Dell et al., 2000).

There is a deep issue here: does the fit of a model to empirical data have any relationship to the support enjoyed by the theory realized by that model? In other words, can fitting patient data help support a theory? And can failing to fit patient data disconfirm a theory? Rapp and Goldrick argue that "It clearly makes no sense to evaluate theoretical claims that make precise quantitative predictions with qualitative behavioral findings. It is equally senseless, however, to evaluate broad theoretical claims with detailed empirical results" (p. 493). Unfortunately, the broad theoretical claim of interactivity does not on its own make precise quantitative predictions. Is it untestable? Is modeling pointless? We would propose no, and rather that broad theoretical claims must eventually demonstrate their ability to be instantiated in models that quantitatively explain data.

The flow of empirical support to broad hypotheses is indirect and gradual. Certainly the failure of a simulation to account for the empirical data does not doom the general theoretical principles used to design it. Only the particular combination of general ideas and specific assumptions embodied in a simulation can be disconfirmed (Ruml & Caramazza, 2000). But at the same time, the general theoretical principles receive only very indirect support if a simulation does match the data. Extensive comparative studies are needed before one can tentatively assign credit or blame to particular assumptions. For example, we have already seen that Dell et al.'s claim that fitting patient data using their interactive model provided support for their theoretical claim of interactivity was a premature judgment. A model without interactivity fits the patient patterns equally well. Even a broad comparative study such as that of Rapp and Goldrick, who examined several different models with different degrees of interactivity, can only begin to suggest theoretical principles that are more or less helpful in constructing models that account for the data under consideration.

Given the diversity of possible modeling paradigms and architectures in psychology, it remains unclear whether we know more now about the necessary (or even plausible) ingredients of the human lexical system than we would have without the many modeling efforts that have been undertaken. If one is willing to make architectural assumptions such as localist representation and activation spreading, then the lexical bias and mixed error effects do seem to necessitate some kind of interactivity.[8] Beyond this obvious conclusion, which requires no support from simulations, we have only limited experience with particular proposals. Interactivity need not even be implemented by feedback, as the recurrent attractor networks of Plaut and Shallice (1993) demonstrate. The conclusions of any modeling study are therefore heavily tied to the myriad of assumptions, architectural and otherwise, that characterize current research. Although modeling is useful and necessary, it is perhaps premature to expect it to contribute strong constraints to cognitive theory. More specifically, if interactivity does not predict patient naming patterns without the help of substantial assumptions, and if enormous work is required to generalize beyond those assumptions, then we would suggest that patient naming patterns are not the right place to look for support for interactivity at the present time.

Having closely examined simulation models of lexical access, we turn now to a consideration of the patient data on their own and what they have to say about the other general theoretical ideas related to interactivity: continuity and globality.

Interactivity, Continuity, and Globality

In Dell et al.'s original paper on the relationship between their model and aphasic patients (1997), they considered three claims: interactivity, continuity, and globality. There is an intuitive explanation for why these three ideas appear together. Imagine an extremely interactive system in which subsystems need input from many other subsystems to correctly compute their own final output. Under such an arrangement, damage in any particular location is likely to have a widespread effect, in proportion to the dependence of other parts of the system on the damaged part. Under such a general degradation of performance, damage anywhere is likely to produce roughly similar output. Following such intuitions, it is not surprising that Dell et al. first considered a theory of damage that did not distinguish particular loci of impairment. And when damage depresses the functioning of all components, it is not surprising that extreme damage in any form would result in random performance. Thus follows the continuity thesis: that extreme damage results in random performance, with intermediate states lying between random and normal behavior.

[8] If one is willing to assume an independent output monitoring system, even this is not certain (Levelt, 1989).

In this rough intuitive sense, high interactivity suggests that even local damage will mimic globality, and globality implies continuity.

The inadequacy of the globality assumption has been discussed extensively elsewhere (Caramazza et al., 2000; Rapp & Goldrick, 2000; Foygel & Dell, 2000; Ruml & Caramazza, 2000). But its close relationship with continuity and interactivity should be noted. For if high interactivity implies globality, and globality is false, then interactivity itself must be tempered. But what of continuity? Is it true? We examine this hypothesis next.

Continuity in Patient Data

Although we have seen that simulations of Dell et al.'s global damage assumption exhibit continuity between random and normal response patterns while the local damage assumptions do not, we have not yet examined whether the patients themselves exhibit continuity. As we mentioned at the beginning of this paper, the continuity thesis states that "more severe aphasic patients have an error pattern that is closer to the error opportunities afforded by the lexicon, whereas less severe aphasic patients have a pattern that is similar to the normal pattern" (Dell et al., 1997, p. 820). By error opportunities, Dell et al. are referring to the error pattern that would result if a patient selected a phonologically legal sequence of phonemes uniformly at random. For English, they estimated that this pattern is dominated by nonwords, unrelated errors, and formal errors, with few semantic and mixed responses. As Dell et al. (2000) further explain, because the continuity thesis limits severe cases to the random pattern, it predicts that "semantic and mixed errors do not increase with the severity of the deficit, while other error categories do" (pp. 636–637). Although Ruml and Caramazza (2000) found little support for the continuity thesis among Dell et al.'s (1997) patients, Dell et al. (2000) note that examining a larger pool of patients could provide a clearer view. Since the continuity thesis is based on Dell et al.'s general theory of lexical access (rather than on the details of any particular implemented computational model), it should hold for our corpus of Italian patients. By checking for a relationship between overall correctness and other error types, we can verify whether lowered correctness is accompanied by movement toward random error rates.

Figure 7 shows the relationships between correctness and nonword, semantic, and mixed errors that occurred among our Italian patients. As in Figures 3– 6, small circles in each panel show each patient's particular combination of correctness and error frequency and shaded triangles indicate regions where patients absolutely cannot lie because the two response frequencies alone would sum to greater than one. Following Dell et al. (1997, figure 4), response frequencies were calculated including the 'other' responses and all patients reported are shown. (Similar results were obtained using data normalized without the 'other' responses.) Panels on the left show patients reported by Dell et al., while panels on the right show our Italian patients. Plus signs mark the performance of English and Italian normals and the random error opportunities computed by Dell et al. for English and by us for Italian (following Dell et al.'s methodology, as we described earlier).

Examining Dell et al.'s data (left panels), we see that the absence of severe cases from their corpus means that there is little evidence regarding the relation between patients with low correctness and the random point. It is difficult to be assured that severely impaired patients with non-random behavior are rare or nonexistent given that Dell et al.'s sample contains so few severely impaired patients. It is those patients that provide a true test of

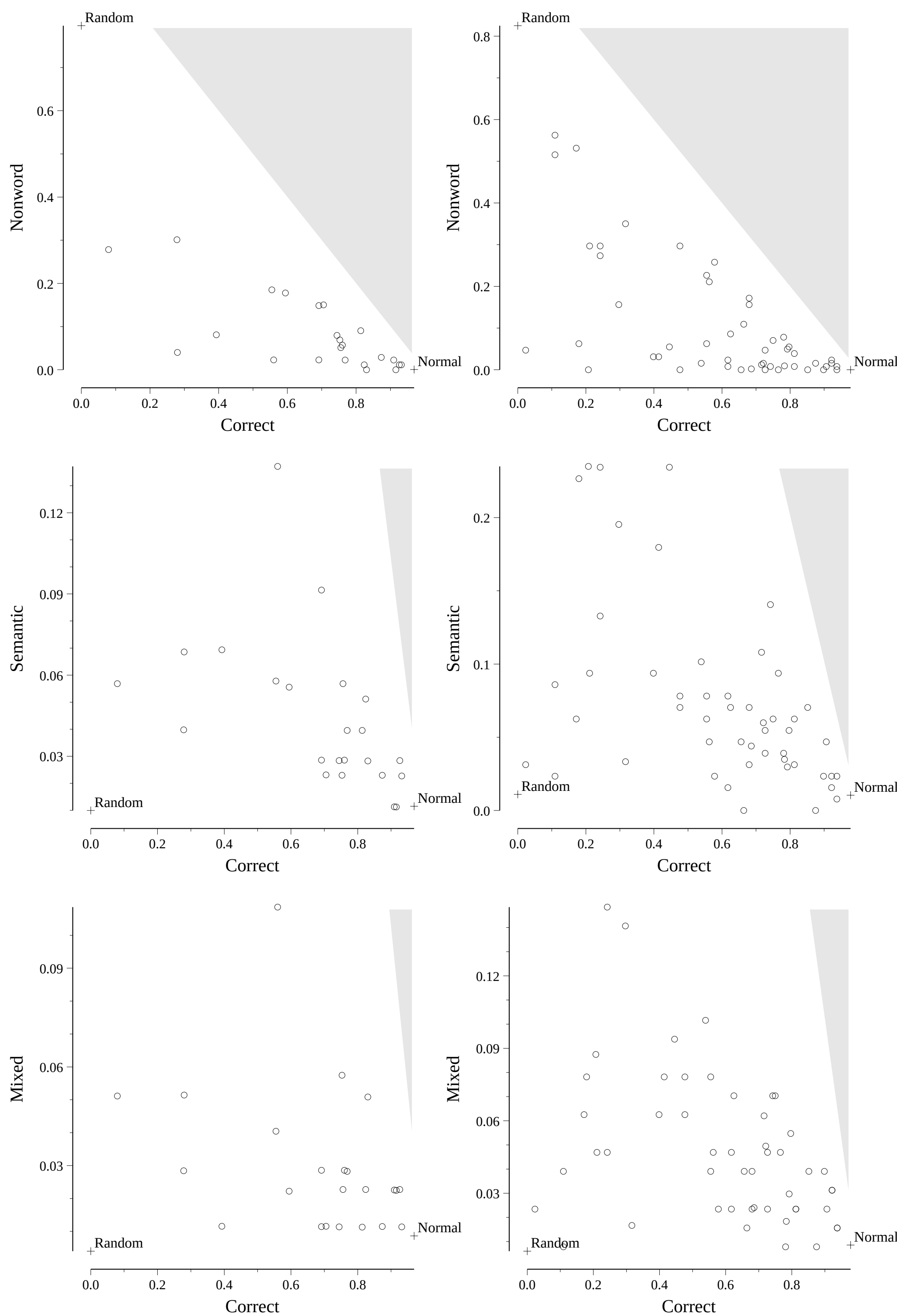

Figure 7. Errors patterns of patients reported by Dell et al. (on the left) and our Italian patients (on the right).

the thesis, since mildly impaired individuals have, by definition, patterns resembling the normal one.

Our Italian corpus, which includes more than twice as many patients as Dell et al.'s sample, does not seem to show strong evidence of the continuity thesis (right panels). On the contrary, we find that patient patterns can be spread quite widely. In the top right panel, we see that several severely impaired patients make very few nonword errors, while continuity predicts that they must make many. (Because this is the response category with the largest variance and the largest difference between normal and random performance, it is presumably the most important for assessing the continuity thesis.) Although of course many impaired patients do make frequent nonword errors, the range of possible patterns shows remarkable diversity. Considering that nonwords are only one possible error response, and hence that patient patterns are very unlikely to be found close to the shaded triangle, the data suggest that nonword errors can be frequent or infrequent, regardless of the severity of the patient's deficit.

Similarly, in the middle and bottom right panels, we see that some severely impaired patients make many semantic or mixed responses, rather than few. The data seem compatible with the view that low correctness merely allows a high percentage of other error frequencies, rather than necessitating convergence to a random error profile. The only way in which these patient patterns could be seen as restricted by the random point is if we were to imagine the continuity thesis as specifying a bounded region that suddenly narrows from allowing the values furthest away from random to occur at only 20–25% correct to insisting on only random values at 0–5% correct. This would not be a substantive prediction at all, as it would exclude patient patterns from only a tiny fraction of the possible space. If continuity is to have any meaning at all, it must place limits on what we can expect to observe. But the data are so widely dispersed from the random point that any meaningful limit would be incompatible with the evidence.

One might suspect that continuity could still be true of patient patterns in a more general way, even while not holding for particular response categories when viewed individually as in Figure 7. To test this idea, we measured how different each patient's complete error pattern was from the random pattern. Figure 8 plots these data as a function of correctness. The difference between each patient's error pattern and the random error pattern was calculated using root mean squared deviation (RMSD), for which a larger value corresponds to a larger difference. Patients reported by Dell et al. are shown at the top panel and our Italian patients are shown in the middle panel. As correctness decreases (toward the left), patient dissimilarity to the random pattern necessarily decreases somewhat, due to the decreased frequency of correct responding. But while some patients with low correctness do exhibit behavior similar to random errors (low RMSD), many other patients do not. Several impaired patients show patterns quite dissimilar from random responding. Patients diverge from normal performance in a trumpet-shaped scatter toward the left of the panels, with both large and small RMSD values at lower correctness. This trend, while only weakly visible in Dell et al.'s data, is clear in our Italian corpus. If the continuity thesis had been correct, we would have observed clustering at low correctness as well as at high correctness, with low correctness patients always exhibiting only small differences from random responding. It appears that the continuity thesis does not hold even when we consider each patient pattern as a whole, allowing individual response types to vary from random

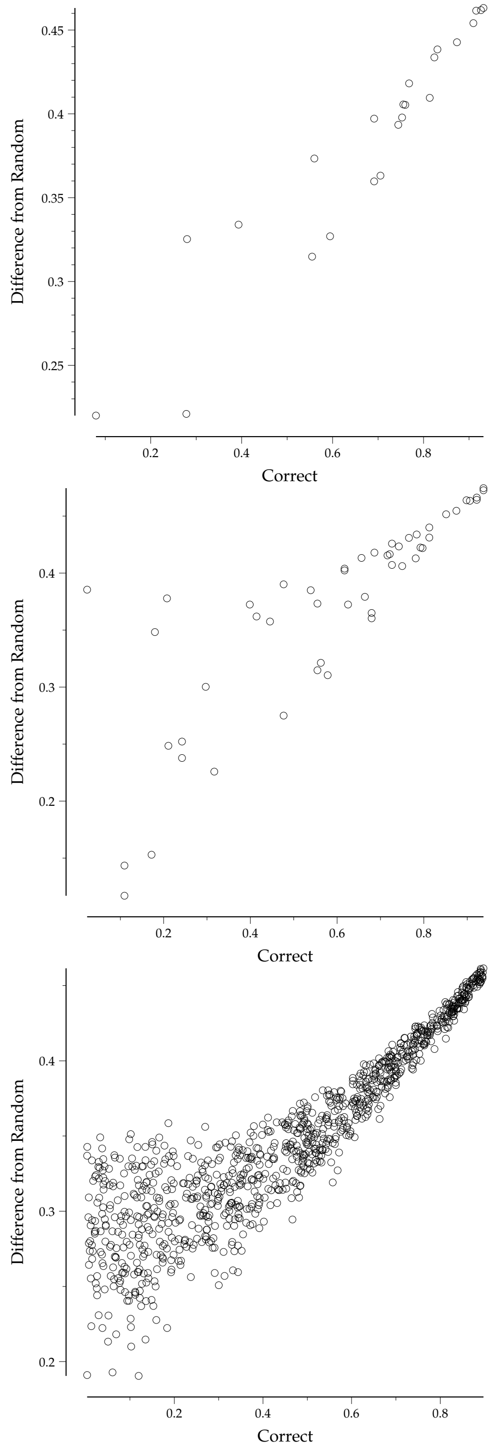

Figure 8. Difference from random responding as a function of correctness for patients reported by Dell et al. (top), our Italian patients (middle), and randomly generated pseudo-patients (bottom).

by averaging over all of them. Instead, it appears that there is no particular trend for low correctness patients to have patterns similar to the random pattern. For comparison purposes, we constructed 1,000 randomly generated 'pseudo-patient' patterns (uniformly distributed and normalized to sum to 1− correct) and measured their difference from the random pattern. These results are shown in the bottom panel. If anything, the randomly generated patterns seem to occupy a narrower range than the patient data, confirming that the patient data is not restricted to any particular trend. We must conclude that patient behavior is not particularly limited by the random error opportunities as severity increases. The only way in which the data can be reconciled with continuity is to gut the thesis of its content.

Dell et al.'s analysis, aside from its use of fewer patients, may have been hampered by its use of error rates averaged across many cases. (Their main evidence for continuity involved plots of group averages for near-normal, moderate-, and high-severity patients.) Clearly, with patients whose naming performance is very good, it is impossible to encounter high error rates. If low correctness merely expands the possible range of error rates, then the average error rate of a high-severity group is certain to be higher than that of a low-severity group, especially in high-variance categories. So finding an average trend toward random opportunities in high-variance categories may merely be an artifact of patient diversity. In a more direct analysis using individual patients from our larger Italian corpus, the continuity thesis finds scant support.

Even though the continuity thesis does not hold for patient data, continuity itself is closely related to several other more conventional ideas. For clarity, we now briefly discuss these connections.

Continuity versus Dissociation

Dell et al.'s continuity thesis is a very strong claim. Dell et al. motivate it as a specific testable instantiation of the conventional idea that aphasic performance can be explained by relatively straightforward modifications of a theory of normal performance. This more general idea has been variously termed the 'transparency assumption' by Caramazza (1992) and Ruml and Caramazza (2000) and 'Continuity' (with a capital C) by Dell et al. (2000). But the continuity thesis is distinct from transparency. Transparency is a standard assumption in cognitive science, whereas continuity predicts random behavior among severely impaired patients. One can easily imagine an alternative to continuity. Suppose that the breakdown of a small component in the lexical system could produce bizarre, severely impaired behavior that is completely at odds with either normal or random performance. This *dissociation thesis* is certainly compatible with transparency but is weaker than the continuity thesis. As Caramazza et al. (2000) note (see also Shallice (1988)), dissociation is weaker because it is compatible not only with behavior resulting from very circumscribed non-random deficits but also with random behavior. Under dissociation, random behavior can be explained as a complex mixture of individual deficits which happen to co-occur. But unlike continuity, dissociation allows for the possibility that these individual deficits could occur in isolation. The continuity thesis is opposed to dissociation because continuity does not allow for behavior that is very different from normal performance and severely impaired but yet highly structured and non-random.

There is an additional idea that connects to both transparency and continuity but is separate from our concerns here. Both continuity and dissociation are compatible with the idea that partial functional damage may lead, through graceful degradation of the neural implementation, to small changes in observed behavior. This would be observed as a kind of mathematical continuity or interpolation among patients: between any two patients one could imagine a third whose damage might be less severe than one and more severe than the other. Like transparency, this interpolation assumption is orthodox in contemporary cognitive science.

Assuming both interpolation and dissociation, one would predict that patients would be observed in a scattered fashion between the various poles defined by normal performance and the complete breakdown of each separate component of the lexical system. The continuity thesis makes the stronger claim that patients will line up between normal and random performance, depending on the amount of damage, with little freedom to include extreme dissociations, especially for severely impaired patients. The close relationship of continuity with the conventional transparency and interpolation assumptions may have obscured its radical departure from the dissociation thesis. Given the many examples in cognitive science of successful models built on the dissociation thesis, it is not surprising that the continuity thesis was not compatible with our patient data.

Other Specific Relatives of Continuity

While the continuity thesis seems untenable, Dell et al. also suggested in passing that variation in patient picture-naming can be summarized in terms other than convergence toward the random point. Because these hypotheses have never been formally evaluated with patient data, we will briefly touch on them here. For instance, Dell et al. (1997) suggest that "a large component of disordered naming can be linked to general severity" (p. 820). This intuitive idea is much weaker than the continuity thesis in that it makes no claims about convergence and makes no specific mention of normal or random behavior. Its plausibility can be verified by examining any corpus of patient error patterns. Figure 2, for instance, clearly shows that it is the frequency of correct responses that varies most among both Dell et al.'s patients and ours.

We can make this claim more formal, and thus testable, by articulating a model that predicts a patient's error profile directly from that patient's correctness. By explicitly testing such a model's VAF (variance accounted for), we can assess the degree to which severity alone accounts for patient behavior. We constructed the simplest possible such model: a set of linear equations that predicts a response category i to be $(\mathit{correct} \times c_i) + m_i$ where the c_i and m_i represent the slopes and intercepts. All 6 response categories (other than correct) were predicted and the result was normalized to sum to 1. We estimated the VAF of the model using cross-validation: for each of the 50 patients we trained the model using least-squares regression on the other 49 patients and then predicted the pattern of the held-out patient from his frequency of correct responses. This leave-one-out technique provides a slightly pessimistic estimate of the model's generalization ability but prevents it from implicitly memorizing its training data. VAF was computed using separate category means and summing over categories, implicitly weighting them by their variances.

Prediction of naming from correctness alone accounted for 71% of the variance in the Italian error profiles. On Dell et al.'s English-speaking patients, the model accounted for 77% of the variance. To gauge this performance, we also tested a principal components

Table 12: Summary of the variance accounted for (VAF) by various simple summary models of patient naming.

	Corpus	
Model	Italian	Dell et al.
Correctness	.71	.77
PCA(1)	.70	.77
PCA(2)	.93	.90
PCA(3)	.98	.93
PCA(4)	.99	.97
Strict continuity	.11	.36
Strict continuity, excluding 'other'	.60	.58

model (PCA) which summarizes a patient's entire pattern into a single number and then attempts to predict the patient's performance from that number. Using the same testing methodology, the PCA model had essentially the same VAF as prediction from correctness. This indicates that not only is correctness the most variable response category across patients but that other categories can be predicted from it according to a simple proportional relationship. One does no better at prediction when using the other categories to form a single numerical summary of a patient.

If we allow PCA to construct multiple summary numbers for a patient error pattern, its predictive accuracy naturally increases. Table 12 shows the VAF of the various models using either correctness alone or PCA with 1 to 4 summary numbers. The results echo those of Ruml and Caramazza (2000), who found that Dell et al.'s patient data (normalized by removing 'other' responses) could be well captured by a two-dimensional linear model. The success of a two- or even three-dimensional model seems surprising, given the variety of patterns that one would expect from the standard dissociation thesis. Remarkably, the Italian data are just as planar, despite the greater variety of patient patterns. Although a grouped score like VAF can be misled by the frequencies of error patterns and obscure important details of individual patient patterns, this analysis confirms one's intuition that accounting for gross patterns in picture naming is not enough on its own to motivate and validate a complex model of language production. Of course, these data still provide an important source of constraints for modelers and theorists, to be used in concert with constraints from other tasks and from linguistics at large.

We also tested the VAF of another variation on the continuity thesis. Dell et al. (1997) claimed that "a considerable portion of the variation among patients can be explained by states intermediate between normality . . . and random responding" (p. 820), although they did not formalize or test this hypothesis specifically. We constructed a model which represents each patient as the closest point that lies directly between normal performance and random behavior. This model had a VAF of 11% for our Italian patients and 36% for Dell et al.'s patients, much less than the first principal component or a model based solely on correctness. Excluding 'other' responses raised the accuracy of this strict continuity hypothesis to 60% (58% for Dell et al.'s patients), which is still less than the VAF of the first principal component for the full data.

We have now considered several ideas closely related to Dell et al.'s continuity thesis. By making these intuitive ideas into explicit formal hypotheses, we were able to evaluate them quantitatively using our corpus of patients. We conclude that continuity, in every form we have considered, is not a particularly helpful notion.

Conclusions

We presented the naming performance of 50 fluent Italian aphasic patients scored using the methodology of Dell et al. (1997). We evaluated the overlap between the patients and three models of lexical access in aphasia, all of which were based on Dell et al.'s interactive theory. Using an automated lexicon construction procedure, we found a model lexicon that matched Italian normal performance and random error opportunities very well. However, none of the three models of aphasic damage we investigated allowed coverage of the patient patterns. The two that eschewed continuity had better coverage. Predictions of patient repetition patterns failed to model most types of errors. In addition, a model similar to that proposed by Foygel and Dell (2000) but completely lacking interactivity did an equally good job of accounting for patient performance. We conclude that aphasic patient data do not lend support to current interactive models of language production.

Examination of the patient data on its own revealed a great variety of error patterns. This diversity seemed incompatible with any substantive interpretation of Dell et al.'s continuity thesis. Error patterns from patients seemed to lie at least as far from random performance as random data. Furthermore, the continuity thesis itself seems directly opposed to the dissociation thesis which has enjoyed widespread success in cognitive science. We also argued that interactivity and continuity are strongly linked to the discredited idea of global damage. We conclude that continuity is false and we are led to speculate that, consonant with the proposal of Rapp and Goldrick (2000), interactivity in lexical access is likely to be limited.

More generally, our investigation demonstrates that fitting empirical data with a computational model does not by itself provide support for the theoretical assumptions of the model. Only extensive efforts to achieve comparable performance with non-interactive models could allow modeling results to support interactivity, for example. Unavoidable assumptions, ranging from the high-level processing architecture to the details of the lexicon, prevent modeling efforts from directly providing constraints on cognitive theory. Although there may be sound reasons for suspecting tight integration of representations involved in language processing, success in modeling patient error patterns is not yet one of them. We believe that modeling is a powerful and useful tool for refining our understanding of theoretical ideas and for connecting to empirical data. Further exploratory work, perhaps along the lines of Rapp and Goldrick's comparative study of the capabilities of various architectures, may lead to important new proposals.

References

Aarts, E., & Lenstra, J. K. (1997). *Local search in combinatorial optimization.* New York: John Wiley & Sons.

Best, W. (1996). When racquets are baskets but baskets are biscuits, where do words come from? A single case study of formal paraphasic errors in aphasia. *Cognitive Neuropsychology*, *13*, 443-480.

Caramazza, A. (1992). Is cognitive neuropsychology possible? *Journal of Cognitive Neuropsychology*, *4*(1), 80–95.

Caramazza, A., Chialant, D., Capasso, R., & Miceli, G. (2000). The representations of vowels and consonants in the brain. *Nature*, *403*, 428–430.

Caramazza, A., Papagno, C., & Ruml, W. (2000). The selective impairment of phonological processing in speech production. *Brain and Language*, *75*(3), 428–450.

Carlesimo, G. A., Caltagirone, C., & Gainotti, G. (1996). The mental deterioration battery: normative data, diagnostic reliability and qualitative analyses of cognitive impairment. *European Neurology*, *36*, 378–384.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589–608.

Cuetos, F., Aguado, G., & Caramazza, A. (2000). Dissociation of semantic and phonological errors in naming. *Brain and Language*, *75*, 451–460.

Dell, G. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Language and Verbal Behavior*, *20*, 611-629.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (2000). The role of computational models in neuropsychological investigations of language: Reply to Ruml and Caramazza. *Psychological Review*, *107*(3), 635–645.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182–216.

Harley, T. A., & MacAndrew, S. B. G. (1995). Interactive models of lexicalisation: Some constraints from speech error, picture naming, and neuropsychological data. In J. P. Levy, D. Bairaktaris, & J. A. Bullinaria (Eds.), *Connectionist models of memory and language* (pp. 311–331). London: UCL Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Martin, N., Dell, G. S., Saffran, E. M., & Schwartz, M. F. (1994). Origins of paraphasias in deep dysphasia: Testing the consequences of a decay impairment to an interactive spreading activation model of lexical retrieval. *Brain and Language*, *47*, 609–660.

Miceli, G., Amitrano, A., Capasso, R., & Caramazza, A. (1996). The remediation of anomia resulting from output lexical damage: Analysis of two cases. *Brain and Language*, *52*, 150–174.

Miceli, G., Benvegnù, B., Capasso, R., & Caramazza, A. (1997). The independence of phonological and orthographic lexical forms: Evidence from aphasia. *Cognitive Neuropsychology*, *14*, 35–70.

Miceli, G., & Capasso, R. (1997). Semantic errors as evidence for the autonomy and the interaction of phonological and orthographic forms. *Language and Cognitive Processes*, *14*, 733-764.

Miceli, G., & Capasso, R. (2001). Word-centred neglect dyslexia: Evidence from a new case. *Neurocase*, *7*, 101-117.

Miceli, G., Capasso, R., & Caramazza, A. (1994). The interaction of lexical and non-lexical mechanisms in reading, writing, and repetition. *Neuropsychologia*, *32*, 317–333.

Miceli, G., Capasso, R., & Caramazza, A. (1999). Sublexical conversion procedures and the interaction of phonological and orthographic lexical forms. *Cognitive Neuropsychology*, *16*, 557-572.

Miceli, G., Capasso, R., Daniele, A., Esposito, T., Magarelli, M., & Tomaiuolo, F. (2000). Selective deficit for people's names following left temporal damage: An impairment of domain-specific knowledge. *Cognitive Neuropsychology*, *17*, 489–516.

Miceli, G., & Caramazza, A. (1988). Dissociation of inflectional and derivational morphology. *Brain and Language*, *35*, 24–65.

Miceli, G., & Caramazza, A. (1993). The assignment of word stress: Evidence from a case of acquired dyslexia. *Cognitive Neuropsychology*, *10*, 273–295.

Miceli, G., Giustolisi, L., & Caramazza, A. (1991). The interaction of lexical and non-lexical processing mechanisms: Evidence from anomia. *Cortex*, *27*, 57–81.

Miceli, G., Laudanna, A., Burani, C., & Capasso, R. (1994). *Batteria per l'analisi dei deficit afasici.* Roma.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377–500.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*(3), 460–499.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*, 107–142.

Ruml, W., & Caramazza, A. (2000). An evaluation of a computational model of lexical access: Comment on Dell et al. (1997). *Psychological Review*, *107*(3), 609–634.

Ruml, W., Caramazza, A., Shelton, J. R., & Chialant, D. (2000). Testing assumptions in computational theories of aphasia. *Journal of Memory and Language*, *43*(2), 217–248.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge University Press.

Shallice, T., & McGill, J. (1978). The origins of mixed errors. In J. Requin (Ed.), *Attention and performance VII* (pp. 193–208). Lawrence Erlbaum Associates.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Weintraub, S., Rubin, N. P., & Mesulam, M. M. (1990). Primary progressive aphasia: Longitudinal course, neuropsychological profile, and language features. *Archives of Neurology*, *47*, 1329–1335.

Appendix A: Patient Data

In this Appendix, we discuss the background of the patients whose performance is discussed in the body of the paper and we present their performance on picture naming and word repetition tasks.

Background

The patients' backgrounds are summarized in Table 13. The etiology of each patient is abbreviated V (vascular), T (traumatic), N (neoplastic), PPA (primary progressive aphasia), HVE (herpes virus encephalitis), SAH (subarachnoid hemorrage), AVM (arteo-venous malformation), MCA (middle cerebral artery) or ICA (internal carotid artery). The lesion site is abbreviated F (frontal), P (parietal), T (temporal), O (occipital), or n/a (not available). The time of testing is given in months post-onset (P.O.). A patient's speech was classified as dysfluent whenever speech was halting or slowed (word-finding pauses were not considered) and dysarthric whenever articulation was slurred or imprecise.

Picture-Naming

Table 14 lists the picture-naming performance of the patients. Patients are sorted in decreasing order of correctness. The number of stimuli is marked as n. The first row for each patient corresponds to our stricter criterion for formal relatedness and the second row scores errors using the method of Dell et al. (1997).

Repetition

In 33 cases, the 128 nouns included in the naming task were also presented for repetition. The examiner pronounced one stimulus at a time, and the subject was asked to repeat it. As in the naming task, the first response produced by the subject was analyzed. 17 subjects were presented with different noun stimuli, and the analyses reported here refer to these stimuli.

Comprehension

The 128 pictures used for naming were also used to evaluate comprehension by means of a word-picture verification task. The examiner presented a picture, while at the same time pronouncing a word, and the subject was asked to say whether or not the word and the picture matched. Each picture was presented three times, in separate sessions (no more than one session per day). In each presentation, it was paired with a different alternative (the correct target, as in pear-"pear"; a semantically related noun, as in pear-"apple"; or, a semantically unrelated noun, as in pear-"dog"). In each session, an approximately equal number of pictures was paired with the correct word, with a semantic foil or with an unrelated foil. The first session was held at least two days after the repetition task. In each session, only one picture-word pair was presented. An item was scored as not having been comprehended when the subject responded incorrectly to one or more of the three word-picture pairs prepared for that item. This task was administered to 40 of the subjects reported here.

In the remaining 10 subjects (AS, SF, ESO, FDI, EMA, GMA, RBO, GIU, FS, CLB), comprehension was assessed by means of a less demanding procedure. These subjects were

Table 13: Patient Background.

Patient	Sex	Age	Edu.	Hand.	Etiology	Lesion site	P.O.	Dysfl.	Dysar.
APA	F	32	13	R	Trauma	left T	44		
AIU	M	43	17	R	V, hemorragic (SAH)	left TP	21		
MPA	F	31	13	R	V, ischemic	left FTP	24		
FVE	M	56	18	L cor	V, hemorragic	right FP	5		
FDP	M	23	8	R	Trauma	left T Pole	24		
GRO	M	51	8	R	V, hemorragic (ruptured AVM)	left F	9		
VFI	F	20	12	R	V, hemorragic (ruptured AVM)	left post T & inf. P	4		
FGU	M	62	13	R	V, ischemic	left TO & R Cerebellum	2		
FSU	F	29	13	R	V, ischemic	Int. capsule & corona radiata	1		
DLI	M	71	8	R	V, hemorragic	left P	1		
DRU	F	50	5	R	V, ischemic	left T ant	16		
AS	F	40	5	amb	V, ischemic	left T & left P	19		
SF	M	46	12	R	Epilepsy	left T polectomy	120		
ESO	M	65	12	R	V, ischemic	left T	2		
SVE	M	70	13	L cor	V, ischemic	right FTP	42	+	
DPA	M	63	13	R	V, ischemic	left P	10		
VFE	M	51	8	R	HVE	left T	32		
PPP	M	65	17	R	V, hemorragic	left T	2		
EPA	M	50	17	R	V, hemorragic	left FTP & basal ganglia	96		
FDI	M	40	17	R	V, hemorragic (ruptured MCA aneurysm)	L insula & corona radiata	6		
EMA	M	66	8	R	V, ischemic	left TP	36		
GMA	M	65	17	R	V, ischemic	left T	12		
IFA	F	52	8	R	V, ischemic	left TP	30		
EMI	M	55	5	R	V, hemorragic (intracerebral hematoma)	left P	4		
PSA	M	50	13	R	V, hemorragic + neoplastic	left T & right F	1		
PDI	F	27	13	R	V, hemorragic (ruptured AVM)	L basal ganglia	11		

(*table continues*)

Table 13: (*Table 13, continued*)

Patient	Sex	Age	Edu.	Hand.	Etiology	Lesion site	P.O.	Dysfl.	Dysar.
MTE	M	75	17	R	V, ischemic	left T	19		
RBO	F	40	13	R	V, hemorragic (ruptured AVM)	left P & left T	18	+	
AME	F	81	12	R	PPA	n/a	6		
GMU	M	47	19	R	V, ischemic (dissection, ICA)	left FP & TP	2		
ECA	M	78	13	R	V, ischemic	left TPO	164		
GIU	M	63	17	R	V, hemorragic (ruptured MCA aneurysm)	left P	50		
GBU	F	58	10	R	PPA	n/a	24	+	
BCO	M	75	17	R	PPA	left T atrophy	36		
GMAR	F	54	8	R	V, hemorragic (ruptured AVM)	left TP	24		
FS	M	73	17	R	V, hemorragic	left T & Insula	209	+	+
ACO	M	27	13	R	HVE	left FTP & right T	9		
RTU	M	67	13	R	HVE	left T, hippocampus & insula	8		
MPU	F	67	5	R	V, ischemic	left P	4		
CLB	M	59	17	R	V, ischemic	left T0	1		
TGU	F	40	11	R	V, hemorragic (ruptured MCA aneurysm)	left FTP	24	+	
MIO	M	52	17	R	V, ischemic	N/A	11	+	+
EGI	F	35	13	R	V, hemorragic (embolisation of AVM)	left TP	36		
CDS	F	40	8	R	V, ischemic	left FTP	3	+	+
PGE	F	41	12	R	HVE	left FT & right T	48		
DFA	M	18	12	R	Trauma	bilateral FTP	8		
WMA	M	48	13	R	V, hemorragic	left FTP	47	+	+
GNI	M	72	15	R	V, ischemic	left posterior	8		
FBI	M	50	18	R	V, hemorragic (intracerebral hematoma)	L lenticular nucleus & int. capsule	26	+	+
RDP	M	60	13	R	HVE	left T	15		

Table 14: The Patients' Picture-Naming Performance Using Both Scoring Criteria.

Patient	n	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Other
APA	128	120	3 (1)	0 (0)	2 (4)	0 (0)	0	3
AIU	128	120	1 (0)	0 (0)	2 (3)	0 (0)	1	4
MPA	128	118	2 (2)	0 (0)	4 (4)	0 (0)	3	1
FVE	128	118	3 (1)	0 (0)	4 (6)	0 (0)	2	1
FDP	128	116	6 (1)	0 (2)	3 (8)	2 (0)	1	0
GRO	128	115	3 (1)	0 (0)	5 (7)	0 (0)	0	5
VFI	128	112	0 (0)	0 (0)	1 (1)	0 (0)	2	13
FGU	128	109	9 (5)	0 (0)	5 (9)	0 (0)	0	5
FSU	128	104	4 (0)	1 (1)	3 (7)	0 (0)	5	11
DLI	128	104	8 (5)	0 (0)	3 (6)	0 (0)	1	12
DRU	128	102	7 (3)	2 (2)	7 (11)	1 (1)	7	2
AS	101	80	3 (3)	1 (1)	3 (3)	1 (1)	5	8
SF	545	427	19 (10)	1 (2)	10 (19)	1 (0)	5	82
ESO	128	100	5 (2)	0 (0)	1 (4)	0 (0)	10	12
SVE	128	98	12 (4)	0 (0)	6 (14)	0 (0)	0	12
DPA	128	96	8 (3)	2 (2)	9 (14)	0 (0)	9	4
VFE	128	95	18 (3)	0 (1)	9 (24)	1 (0)	1	4
PPP	128	93	7 (2)	4 (4)	6 (11)	0 (0)	6	12
EPA	128	93	5 (4)	0 (0)	3 (4)	0 (0)	0	27
FDI	384	277	23 (11)	1 (3)	19 (31)	2 (0)	6	56
EMA	500	358	54 (19)	5 (6)	31 (65)	1 (1)	6	45
GMA	500	343	22 (9)	0 (1)	12 (25)	1 (0)	1	121
IFA	128	87	4 (2)	1 (1)	3 (5)	1 (1)	22	10
EMI	128	87	9 (6)	0 (0)	5 (8)	0 (0)	20	7

(*table continues*)

submitted to a 40-item word-picture matching task included in the screening procedure for aphasia. For these tasks, the examiner said a word aloud or presented a written word, while showing two pictures. In both tasks, a picture always corresponded to the stimulus, whereas the other represented in 20 cases a semantic alternative (e.g., fork and spoon for the stimulus word "fork"), and in 20 cases a phonological/orthographic alternative (e.g., book and cook for the stimulus word "book"). The number of incorrect responses produced to the semantic and to the phonemic foils is provided for these subjects. Comprehension data for all subjects are reported in Table 16.

Table 14: (*Table 14, continued*)

PSA	128	85	0 (0)	2 (2)	2 (2)	0 (0)	14	25
PDI	128	84	6 (5)	0 (0)	5 (6)	0 (0)	0	33
MTE	128	80	9 (2)	3 (3)	9 (16)	1 (1)	11	15
RBO	128	79	2 (0)	0 (0)	3 (5)	0 (0)	3	41
AME	128	79	10 (5)	0 (0)	6 (11)	0 (0)	1	32
GMU	128	74	3 (1)	7 (7)	3 (5)	0 (0)	33	8
ECA	128	72	6 (2)	3 (3)	6 (10)	0 (0)	27	14
GIU	128	71	10 (4)	1 (1)	10 (16)	0 (0)	29	7
GBU	128	71	8 (2)	0 (1)	5 (11)	2 (1)	8	34
BCO	128	69	13 (4)	0 (0)	13 (22)	0 (0)	2	31
GMAR	128	61	9 (4)	0 (1)	10 (15)	1 (0)	0	47
FS	128	61	10 (3)	3 (4)	8 (15)	1 (0)	38	7
ACO	128	57	30 (13)	0 (0)	12 (29)	0 (0)	7	22
RTU	128	53	23 (12)	2 (4)	10 (21)	2 (0)	4	34
MPU	128	51	12 (7)	1 (2)	8 (13)	1 (0)	4	51
CLB	60	19	2 (0)	3 (6)	1 (3)	4 (1)	21	10
TGU	128	38	25 (11)	2 (3)	18 (32)	2 (1)	20	23
MIO	128	31	17 (5)	3 (7)	6 (18)	8 (4)	38	25
EGI	128	31	30 (9)	3 (4)	19 (40)	3 (2)	35	7
CDS	128	27	12 (7)	6 (8)	6 (10)	1 (0)	38	38
PGE	183	38	43 (16)	1 (2)	16 (43)	2 (1)	0	83
DFA	128	23	29 (12)	4 (5)	10 (27)	1 (0)	8	53
WMA	128	22	8 (5)	1 (1)	8 (11)	0 (0)	68	21
GNI	128	14	3 (1)	18 (22)	1 (3)	12 (8)	72	8
FBI	128	14	11 (5)	6 (8)	5 (11)	6 (4)	66	20
RDP	128	3	4 (1)	6 (15)	3 (6)	22 (13)	6	84

Appendix B: Constructing a Model Lexicon

Given desired error opportunities and phonological overlap values, it is not obvious how to construct a lexicon with those characteristics. Changing the phonological and semantic relatedness of words to match the desired opportunities must be done carefully in order to preserve the desired phonological overlap between and within semantic categories. Rather than attempt laborious hand-tuning, we developed an automated procedure which tries to find a suitable lexicon. This appendix describes the procedure and the criteria it uses to judge potential lexicons.

Our method is a simple multistart hill-climbing search. Such methods have become popular in the operations research and artificial intelligence communities for solving a wide variety of combinatorial optimization problems (Aarts & Lenstra, 1997). Our procedure starts with a random initial lexicon and then chooses a modification at random in the hope of finding a variation on the current lexicon that matches the desired characteristics better. If an improvement is found, that lexicon replaces the current one. If 1,000 consecutive

Table 15: The Patients' Repetition Performance Using Both Scoring Criteria.

Patient	n	Corr.	Sem.	For.	Mixed	Unrel.	Non.	Other
APA	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
AIU	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
MPA	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
FVE	128	125	0 (0)	1 (1)	0 (0)	0 (0)	2	0
FDP	15	15	0 (0)	0 (0)	0 (0)	0 (0)	0	0
GRO	15	15	0 (0)	0 (0)	0 (0)	0 (0)	0	0
VFI	15	15	0 (0)	0 (0)	0 (0)	0 (0)	0	0
FGU	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
FSU	128	122	0 (0)	2 (2)	0 (0)	0 (0)	3	1
DLI	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
DRU	128	119	1 (1)	1 (1)	0 (0)	0 (0)	7	0
AS	290	255	0 (0)	3 (3)	0 (0)	0 (0)	27	5
SF	15	15	0 (0)	0 (0)	0 (0)	0 (0)	0	0
ESO	15	11	0 (0)	0 (0)	0 (0)	0 (0)	3	1
SVE	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
DPA	128	117	0 (0)	0 (0)	0 (0)	0 (0)	11	0
VFE	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
PPP	15	13	0 (0)	0 (0)	0 (0)	0 (0)	2	0
EPA	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
FDI	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
EMA	195	189	0 (0)	3 (3)	0 (0)	0 (0)	3	0
GMA	15	15	0 (0)	0 (0)	0 (0)	0 (0)	0	0
IFA	27	22	0 (0)	2 (2)	0 (0)	0 (0)	3	0
EMI	128	104	0 (0)	1 (1)	0 (0)	0 (0)	23	0

(*table continues*)

attempts at modification fail to find any improvements, the procedure starts with a new random lexicon. We always remember the best lexicon seen. The quality of any candidate lexicon can be quantified numerically by averaging its matches on our two criteria. We found this simple hill-climbing strategy to be surprisingly effective. In contrast, experiments in which we merely generated many random lexicons from scratch (rather than by modification from a current candidate) proved fruitless.

We found that the precise formulation of the quality measure for the lexicons significantly influenced which lexicons were found. After several pilot experiments, we settled on an *ad hoc* measure based on X^2. (The quantity we calculate has no intuitive probabilistic interpretation as in a χ^2 test; we mere use the same formula as one would use to calculate the X^2 statistic.) Each candidate lexicon was evaluated using 10 terms: the 6 response frequencies, the 2 phonological overlaps (within and between semantic categories), whether or not the within-category overlap was greater than across (represented as 1 if greater, as desired, and 0 otherwise), and the number of opportunities within {semantic, formal, mixed,

Table 15: (*Table 15, continued*)

PSA	88	87	0 (0)	1 (1)	0 (0)	0 (0)	0	0
PDI	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
MTE	128	126	0 (0)	0 (0)	0 (0)	0 (0)	2	0
RBO	15	14	0 (0)	0 (0)	0 (0)	0 (0)	1	0
AME	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
GMU	128	120	0 (0)	0 (0)	0 (0)	0 (0)	7	1
ECA	128	93	7 (2)	3 (4)	3 (8)	1 (0)	17	4
GIU	128	98	0 (0)	1 (1)	0 (0)	0 (0)	29	0
GBU	15	12	0 (0)	0 (0)	0 (0)	0 (0)	1	2
BCO	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
GMAR	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
FS	195	151	0 (0)	9 (9)	2 (2)	0 (0)	30	3
ACO	128	91	0 (0)	3 (3)	1 (1)	0 (0)	33	0
RTU	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
MPU	15	13	0 (0)	1 (1)	0 (0)	0 (0)	1	0
CLB	30	1	0 (0)	4 (5)	1 (1)	1 (0)	5	18
TGU	128	111	0 (0)	1 (1)	1 (1)	0 (0)	15	0
MIO	128	73	0 (0)	3 (3)	1 (1)	0 (0)	51	0
EGI	15	6	0 (0)	0 (0)	0 (0)	0 (0)	7	2
CDS	128	92	0 (0)	2 (2)	1 (1)	0 (0)	31	2
PGE	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
DFA	128	128	0 (0)	0 (0)	0 (0)	0 (0)	0	0
WMA	128	45	0 (0)	2 (2)	1 (1)	0 (0)	79	1
GNI	15	0	0 (0)	0 (0)	0 (0)	0 (0)	5	10
FBI	15	6	0 (0)	0 (0)	0 (0)	0 (0)	8	1
RDP	15	15	0 (0)	0 (0)	0 (0)	0 (0)	0	0

and unrelated} that were greater than zero (all 4 were desired positive). These 10 terms were compared with the desired set of terms using X^2. The last two terms helped the search procedure quickly identify promising lexicons, while the others promoted fine-tuning. The phonological overlap terms were divided by 5 to reduce their effect on the evaluation metric relative to the error opportunities.

The generation of initial lexicons and the modification operations used a biased random process intended to encourage useful modifications. For instance, changing a word's pronunciation was attempted 12 times as often as the addition of a new word. To ensure a realistic phonemic inventory and prevent mistakes such as the creation of a lexicon with more vowels than consonants, Italian consonants and vowels were used during the search. Testing for formal relatedness was done using the phonemes, as it would be for patient behavior, rather than using the nodes of the network, which are marked for order.

Using these techniques, the multistart hill-climbing search found many reasonable lexicons. The procedure tended to gradually increase the number of the words in the lexicon during the search. Using more words allowed the random opportunity for a correct

Table 16: The Patients' Comprehension Performance.

Patient	n	% Correct
APA	128	1.000
AIU	128	.992
MPA	128	.969
FVE	128	.977
FDP	128	.977
GRO	128	1.000
VFI	128	.992
FGU	128	.969
FSU	128	.992
DLI	128	.977
DRU	128	.883
AS	40	.950
SF	40	.950
ESO	40	.975
SVE	128	.906
DPA	128	.922
VFE	128	.734
PPP	128	.844
EPA	128	.937
FDI	40	1.000
EMA	40	.950
GMA	40	.975
IFA	128	.922
EMI	128	.961

(*table continues*)

response to decrease and often allowed a closer match to the desired opportunities. The lexicon used in the body of the paper was manually chosen from among those found by the search procedure in an attempt to retain accuracy using the smallest possible number of words. The search algorithm is quite general and has also proved useful in obtaining a model lexicon for English.

Table 16: (*Table 16, continued*)

PSA	128	.977
PDI	128	.961
MTE	128	.953
RBO	40	.950
AME	128	.805
GMU	128	.977
ECA	128	.750
GIU	40	.950
GBU	128	.867
BCO	128	.797
GMAR	128	.820
FS	40	.875
ACO	128	.516
RTU	128	.578
MPU	128	.891
CLB	80	.913
TGU	128	.797
MIO	128	.781
EGI	128	.594
CDS	128	.875
PGE	128	.547
DFA	128	.633
WMA	128	.781
GNI	128	.883
FBI	128	.648
RDP	128	.344